2019

A Data Mining Framework for Improving Student Outcomes on Step 1 of the United States Medical Licensing Examination

James Clark
Nova Southeastern University, jimmy_clark@yahoo.com

A Data Mining Framework for Improving Student Outcomes on Step 1 of the United States Medical Licensing Examination

by

James P. Clark

A dissertation submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy
in
Information Systems

College of Engineering and Computing
Nova Southeastern University

2019

We hereby certify that this dissertation, submitted by James Clark, conforms to acceptable standards and is fully adequate in scope and quality to fulfill the dissertation requirements for the degree of Doctor of Philosophy.

Steven R.Terrell, Ph.D.
Chairperson of Dissertation Committee

3/11/19
Date

Martha M. Snyder, Ph.D.
Dissertation Committee Member

3/11/19
Date

Ling Wang, Ph.D.
Dissertation Committee Member

3/11/19
Date

Approved:

Meline Kevorkian, Ed.D.
Interim Dean, College of Engineering and Computing

3/11/19
Date

College of Engineering and Computing
Nova Southeastern University

2019

An Abstract of a Dissertation submitted to Nova Southeastern University
in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy

A Data Mining Framework for Improving Student Outcomes on Step 1 of the United States Medical Licensing Examination

by

James P. Clark

2019

Identifying the factors associated with medical students who fail Step 1 of the United States Medical Licensing Examination (USMLE) has been a focus of investigation for many years. Some researchers believe lower scores on the Medical Colleges Admissions Test (MCAT) are the sole factor used to identify failure. Other researchers believe lower course outcomes during the first two years of medical training are better indicators of failure. Yet, there are medical students who fail Step 1 of the USMLE who enter medical school with high MCAT scores, and conversely medical students with lower academic credentials who are expected to have difficulty passing Step 1 but pass on the first attempt. Researchers have attempted to find the factors associated with Step 1 outcomes; however, there are two problems associated with their methods used. First is the small sample size due to the high national pass rate of Step 1. And second, research using multivariate regression models indicate correlates of Step 1 but does not predict individual student performance.

This study used data mining methods to create models which predict medical students at risk of failing Step 1 of the USMLE. Predictor variables include those available to admissions committees at application time, and final grades in courses taken during the preclinical years of medical education. Models were trained, tested, and validated using a stepwise approach, adding predictor variables in the order of courses taken to identify the point during the medical education continuum which best predicts students who will fail Step 1. Oversampling techniques were employed to resolve the problem of small sample sizes. Results of this study suggest at risk medical students can be identified as early as the end of the first term during the first year. The approach used in this study can serve as a framework which if implemented at other U.S. allopathic medical schools can identify students in time for appropriate interventions to impact Step 1 outcomes.

Acknowledgements

I would like to express my deepest appreciation to my advisor and committee chair Dr. Steven Terrell for his wisdom and support throughout the dissertation process, and to committee members Dr. Martha Snyder, and Dr. Ling Wang for their support and contributions to this completed report. I am grateful for your dedication and commitment to excellence in the pursuit of scholarly activities, and your willingness to share this passion with your students.

I also wish to thank members of the Baylor College of Medicine community who supported me during my dissertation journey, specifically Dr. Alicia Monroe, Dr. Jennifer Christner, and Mr. Lee Lieber.

And finally, a huge thank you to my husband Dr. Matthew Thompson for his support and encouragement, and to my family and friends who helped along the way.

Table of Contents

Appendices

List of Tables

List of Figures

Chapter 1

Introduction

Background

The Association of American Medical Colleges (AAMC) predicts a shortage of 121,300 physicians in the United States by 2030. This estimate is driven largely by the expected growth in the population, specifically an increase in those age 65 and older who have higher demands for medical care, and an increased demand in underserved populations (Dall, West, Chakrabarti, Reynolds & Iacobucci, 2018). Since medical training can take up to ten years, medical schools are under increasing pressure to fill the gap; however, capacity at U.S. medical schools is limited and the admissions process is highly competitive. Admissions committees select and matriculate students who are most likely to complete the full medical school curriculum, pass required board examinations, and continue to residency (Gay, Santen, Mangrulkar, Sisson, Ross & Zaidi, 2018).

There is not one prescribed application process for U.S. medical schools, but many schools use a centralized application service provided for member institutions of the AAMC. An integral part of the application process is the Medical Colleges Admissions Test (MCAT), used for over 80 years, which allows admissions committees to evaluate applicants' knowledge and skills needed to be successful in medical school. In addition to MCAT scores, applicants provide information about themselves, their preparation for medical school by way of coursework taken in undergraduate or post-baccalaureate programs, personal statements, and letters of recommendation

("Navigate your journey," 2018). Admission committee members review each applicant's qualifications based on information provided during the application process and letters of recommendation to determine those invited for campus interviews. It is during campus interviews when applicants have an opportunity to learn more about the medical school and interviewers can assess qualities necessary to be good physicians, such as "compassion and empathy, personal maturity, oral communication skills, service orientation, and professionalism" (Monroe, Quinn, Samuelson, Dunleavy, & Dowd, 2013, p. 675).

If medical student selection is based solely on cognitive aptitude, exceeding minimally acceptable undergraduate grade point averages (UGPA) and scores from the MCAT would separate those who have access to medical education, and a career as a physician, from those who do not. Instead, the AAMC urges medical schools to review applicants holistically, with a balanced consideration of UGPA, MCAT scores, and other attributes and life experiences when making acceptance decisions ("Holistic Review," 2018). However, because medical training is cognitively challenging, admissions committees accept and matriculate students who are able to withstand the rigor of medical school and pass licensure examinations. In the past, higher MCAT scores and UGPA have been used as indicators of this ability, but with holistic review, there are applicants who matriculate with lower MCAT scores and UGPA, who are just as likely to be successful in medical school (Monroe et al., 2013; Sesate, Milem, McIntosh, & Bryan, 2017).

The structure of medical education can vary by school, but a typical curriculum consists of preclinical and clinical years, each lasting two years for a total of four years of

medical training. The preclinical part of the curriculum focuses on the foundational sciences necessary to practice medicine, and is mostly lecture-based, with case studies in smaller groups, and laboratory experiences. The clinical years are more commonly known by their clinical rotations, with students working directly with physicians and other trainees in clinical settings observing patient interactions. High-stakes standardized examinations are often used for students to progress though the medical curriculum, and in some cases serve as practice opportunities for medical licensure examinations (Dezee, Artino, Elnicki, Hemmer, & Durning, 2012).

One indicator of academic success in medical school is performance on Step 1 of the United States Medical Licensing Examination (USMLE), a series of three high-stakes examinations, called Steps, which is required for medical licensure, and for the unsupervised practice of medicine. Because Step 1 assesses the knowledge of the foundational sciences necessary to the practice of medicine, a typical series of courses during the preclinical years of medical training, it is conceivable that 100% of students may pass; however, that is not the expectation, and historically has not been the case. According to recent USMLE performance data, the national failure rate of medical students taking Step 1 for the first time in 2017 was 4% (USMLE Performance Data, 2018). National failure rates may be low; however, some medical schools report failure rates up to 15% (Schwartz, Lineberry, Park, Kamin, & Hyderi, 2018).

Passing Step 1 is important for three reasons. First, it is often a requirement for promotion to later years of medical training, and a requirement for graduation. Second, passing Step 1 is the first step in the pathway to medical licensure, and required to qualify for subsequent step examinations. Third, passing Step 1 has been cited as the most

important factor when selecting applicants for residency program interviews (National Resident Matching Program, 2018). Many students consider Step 1 to be a pass or fail examination; however, the majority of residency programs consider Step 1 scores for acceptance, especially for competitive programs such as dermatology and orthopedic surgery.

While not directly responsible for Step 1 outcomes, medical schools have a responsibility to monitor student progress, and to offer assistance when necessary in an attempt to improve outcomes. Recent reports have confirmed the ability of MCAT scores alone to reliably predict Step 1 outcomes (Glaros, Hanson, & Adkison, 2014; Burns & Garrett, 2015); but because of holistic review practices during the admissions process, applicants with lower MCAT scores are accepted, putting them at risk for failing Step 1.

Problem Statement

Step 1 failures have been attributed to lower MCAT scores (Gauer et al., 2016), and lower grades in preclinical courses (Sesate et al., 2017). There are also students who enter medical school with higher MCAT scores but fail Step 1 for no apparent reason (Sesate et al., 2017), and students who enter with lower MCAT scores who pass Step 1 (Monroe et al., 2013). Researchers have attempted to find the factors associated with Step 1 outcomes; however, there are two problems associated with their methods used. First, the smaller sample size due to the low failure rate of step 1 makes it difficult to predict performance (Kleshinski et al., 2009), and is a possible cause for recent research finding no correlation between MCAT scores and Step 1 outcomes (Giordano et al., 2016). Second, research using multivariate regression models indicate correlates of Step 1 performance (Hu et al., 2016; Lee et al., 2017), but does not predict student

performance, leaving the unanswered question "what does it look like for individual students" (Lee et al., 2017, p. 6).

Increasing the sample size to include medical students who scored slightly above the minimum passing score, in addition to medical students who failed, may improve the ability to determine the factors needed to predict Step 1 performance (Hu et al., 2016). Additionally, employing methods other than multivariate regression models may help medical school administrators identify students at risk of failing Step 1, as recommended by Lee et al. (2017). Holistic review admission practices have increased the diversity of applicants accepted into medical school (Monroe, et al., 2013); but identifying medical students who risk failing Step 1, regardless of MCAT scores, may help improve Step 1 outcomes. For this study, an at-risk medical student is any student who risks failing Step 1. Using a lower MCAT score and UGPA as indicators of an at-risk student will ignore those students entering medical school with higher UGPA and MCAT scores who fail Step 1 for no apparent reason (Sesate et al., 2017), or have difficulty during their first two years of medical school, but will not seek assistance (Winston et al., 2014).

Dissertation Goal

The purpose of this study was to identify the factors related to Step 1 failure, and to identify the medical students at risk of failure without using MCAT scores or UGPA as sole indicators. This study adds to the current knowledge of Step 1 outcomes by addressing the deficiencies identified from prior research in this area: the small sample size because of the low Step 1 failure rate (Kleshinski et al., 2009; Giordano et al., 2016), and the ability to predict Step 1 outcomes at the student level, a problem of recent research noted by Lee et al. (2017). To address the sample size deficiency, a wider net

was cast to identify students at risk of failing Step 1: students who failed Step 1, and students who passed Step 1, but within one standard deviation of the mean score, as suggested by Hu et al. (2017). This will be described in greater detail in the third chapter.

Predictive modeling using data mining methods was used to identify the at-risk medical students. Chen and Fawcett (2016) define data mining as the use of computational techniques to identify relationships in large sets of data or predict what is likely to occur given a certain scenario. Clow (2013) applies the concept of predictive modeling to education by developing a model "which produces estimates of likely outcomes, which are then used to inform interventions designed to improve these outcomes" (p. 688). In this study, the set of data includes factors available during the admissions process (i.e. demographics, MCAT scores, UGPA), and course outcomes from the preclinical years of medical school.

Prediction models have been used in business to predict customer churn (Lee, Kim, & Lee, 2017), to find new customers using social networks (Zhao, King, Lye, Zeng, & Yuan, 2017). In medical research, predictive models have been used to aid in clinical decision making (Chen & Fawcett, 2016), and to improve cardiovascular care (Rumsfeld, Joynt, & Maddox, 2016). Predictive modeling is relatively new in education but has recently been used to predict high school dropouts (Marquez-Vera, Dano, Romero, Noaman, Fardoun, & Ventura, 2016), in higher education to predict freshman student attrition (Thammasiri, Delen, Meesad, & Kasap, 2014), and university course performance (Kostopoulos, Lipitakis, Kotsiantis, & Gravvanis, 2017). However, all three examples underscore a problem prevalent in predictive models applied to education problems, that is the case of the outcome in question (e. g. high school dropouts,

freshmen attrition, course outcomes) is not evenly balanced between students who exhibit the outcome and those who do not. If attempting to predict high school dropouts or freshmen attrition, there are far less students who dropout than those who do not dropout. Similarly, for course outcomes there is a larger majority of students who pass a course than the smaller group, often called the majority and minority classes respectively. For each of these studies, predictive model accuracy can be misleading because the majority class contributes much more to overall accuracy then the minority class (Marquez-Vera et al., 2016). Based on prior use in educational settings, there is an opportunity to extend predictive modeling to a medical school for Step 1 student outcome prediction; however, because of the 4% national Step 1 failure rate, the imbalance between medical students who pass Step 1 and those who fail must be addressed.

Research Questions

This study was guided by the following research questions:

- Research Question 1: What are the factors associated with Step 1 failures?
- Research Question 2: Can data mining algorithms be used to identify medical students at risk of Step 1 failure?
- Research Question 3: How does the expected imbalance between students passing Step 1 and those failing Step 1 impact the use of data mining algorithms to identify students at risk of Step 1 failure?

Research question 1 is an examination of preadmissions variables, those available during the admissions process, and curricular measures, the outcomes from courses during the preclinical years of medical school, to identify the factors associated with Step 1 failure. Of interest is the relationship between MCAT scores and Step 1 outcomes.

Recently, Giordano et al. (2016) found no correlation between these measures, which contradicts much of the research literature.

Research question 2 is an attempt to answer the question posed by Lee and colleagues (2017) when they asked what Step 1 outcomes look like for individual students, especially those who entered medical school with higher MCAT scores, but failed Step 1 for no apparent reason (Sesate et al., 2017), and students who entered with lower MCAT scores who pass Step 1 (Monroe et al., 2013). Research question 3 requires special consideration to the expected imbalance between students who passed Step 1, the majority group, and students who failed Step 1, the minority group (Abeysinghe, Hung, Bechikh, Wang, & Rattani, 2018).

Predictive models which identify students at-risk of failing Step 1 can be used by admissions committees during the holistic applicant review process to identify the academically most capable, and those who show promise to be good physicians, but will need assistance to succeed, as suggested by Monroe et al. (2013). Additionally, factors associated with Step 1 failure which include curricular measures can be used to inform decisions made by committees which evaluate student progress and recommend advancement and promotion during medical training.

Relevance and Significance

Prior to holistic review practices for medical school admissions, it was common practice for a school to establish a minimum UGPA and MCAT score, eliminating all applicants who fell below the minimum, including applicants considered underrepresented in medicine (URM) with MCAT and UGPA below non-URM students. Holistic review practices allow medical schools to create diverse classes of students able

to meet the medical needs of a growing diverse population, an important requirement to achieving health equity (Elks et al., 2018). Admissions decisions made solely on UGPA and MCAT could eliminate applicants who possess other qualities needed to become a good physician. Holistic review has changed how admissions committees evaluate applicants, giving a new perspective to UGPA and MCAT scores (Capers et al., 2018).

UGPA and MCAT scores may serve as a guideline for identifying students at-risk of Step 1 failure but can no longer be the only factors. The goal is to identify at-risk students and to offer support programs designed to improve medical school outcomes. Medical schools use a variety of methods to offer support to students before and during medical training. Programs offered prior to the start of the first year prepare students for the rigor of medical training and provide strategies to improve study habits. Program participants are typically selected based on lower MCAT scores and UGPA, and in some cases ethnicity, gender, or age.

Segal, Giordani, Gillum, and Johnson (1999) described a program at the University of Michigan School of Medicine designed to assist students recover from academic difficulties, reporting a 93% improvement in medical school outcomes. Other medical schools have since reported similar outcome improvements when implementing academic support programs (Lieberman et al., 2008; Glaros et al, 2014; Winston et al., 2014); however, in many cases, MCAT scores were used as the sole identifying factor of at-risk medical students.

Not every academic support program achieves the desired results. Hairrell, Smith, McIntosh, and Chico (2016) described a program offered to at-risk medical students before the first year begins. The program was designed to prepare selected

students for the rigor of medical training, and to provide resources to improve study skills. Students were selected based on lower MCAT and UGPA, URM (non-white ethnicity) and older students. There was not a significant difference in the grades between those who attended the program and those who did not, but participants did report an increase in a sense of belonging and confidence. Heck et al. (2017) surveyed 116 medical schools in the U.S. to determine the prevalence of pre-matriculation programs. One quarter of the medical schools responding reported the use of a student assistance program before the first year begins. Many of the schools allow participation by any incoming medical student; however, half of the schools give special consideration to students underrepresented in medicine, and students with lower UGPA and MCAT scores. The majority of participants in these programs did graduate on time. Schneid et al. (2018) described a program at the University of California San Diego School of Medicine offered to all admitted students, but students with lower UGPA, MCAT scores, or underrepresented in medicine are encouraged to attend by the dean. Performance in the program was found to correlate with preclinical course outcomes, but not Step 1.

Barriers and Limitations

Creating a model which accurately predicts medical student outcomes has been a goal in medical education for many years (Lee et al., 2017). This is a goal of the current study, but there are barriers present. The first barrier is related to data access for model creation. There are research datasets available from the AAMC, but they would only have data provided by applicants, but not course outcomes or results of Step exams. USMLE outcomes could be provided by the National Board of Medical Examiners (NBME), but they do not have application variables, or a way to join datasets between

the two. The only option is requesting data from one medical school, but this presents problems with generalization of the results; however, the data mining framework used to create the model in this study could be used at other medical schools seeking similar results.

Conclusions drawn from the findings of this study will only be relevant to the medical school participating in the study. This is a limitation noted by Schwartz et al. (2018) when they studied the effect of a student-initiated study program on Step 1 outcomes, commenting that other medical schools are likely to find different results because of the unique nature of their mission-driven admission policies. This is not uncommon since admissions processes and requirements can vary across medical schools. Even with a holistic review process of applicants, medical school admissions committees will select applicants based on the unique mission of the school, potentially eliminating other equally qualified applicants (Ellaway, Malhi, Baja, Walker & Myhre, 2018). For example, if a mission of a Texas medical school is to prepare primary care doctors who desire to care for underserved populations in Texas, priority consideration will be given to applicants with similar interests.

The medical school curriculum during the preclinical years of training can vary across medical schools. In many U.S. medical schools, the first two years of training is considered the preclinical years, or the years before clinical rotations, when medical students are taught the basic science fundamentals necessary for the practice of medicine. However, there is not one prescribed curriculum for all schools for the preclinical years. Additionally, some medical schools have decided to reduce the time of the preclinical years to 18 months, allowing students a lengthier period to observe patients in a clinical

setting.

Beginning in 2015, applicants to medical schools began taking a revised version of the MCAT, which included a change in the content tested, the number of questions, and the structure of the scores. The AAMC, who administers the MCAT, offers guidance to admissions committees with respect to score interpretation, but does not suggest a valid correlation between old and new scores. According to the AAMC, the scores are not comparable because the new MCAT tests different things, contributing to the intentional change in the structure of the scores ("About the MCAT Exam", n.d.). This threat can be controlled by limiting the sample to students who took the MCAT before or after the 2015 change. Future research will be needed to include the new MCAT in prediction models when more Step 1 outcome data is available.

Medical student self-directed study behaviors have not been included in this study but will be noted as a need for additional research. Increased study time, usage of review books, and attempts at more practice questions have been associated with higher Step 1 outcomes (Burk-Rafel, Santen & Purkiss, 2017) but not available in student information systems. Surveys can be created to capture this information to be used in future prediction models.

Delimitations

There are two types of medical schools in the United States. Allopathic schools grant Medical Doctor degrees, require the MCAT for admissions, and the USMLE for medical licensure. Not all allopathic medical schools in Canada require applicants to take the MCAT. Osteopathic medical schools grant Doctor of Osteopathic Medicine degrees, require the MCAT for admissions, but students have a choice of taking the USMLE or

the Comprehensive Osteopathic Medical Licensing Examination (COMLEX) for medical licensure. This study will also be delimited by the medical school participating in this study, which will be described in the methodology chapter.

For this study, all references to *medical school* will indicate allopathic, MD-degree granting schools, in the Unites States because they have similar admission requirements, follow a similar curricular format of a preclinical block for the study of foundational sciences followed by a clinical block, and have similar licensure and graduation requirements. Similarly, *preclinical* years will be used to refer to the portion of the medical school curriculum devoted to basic sciences courses, and *clinical* years will refer to the time spent in clinical rotations, unless reference to a specific year of medical training is needed.

Definition of Terms

Algorithm: Instructions used to systematically transform input to output (Alpaydin, 2010).

Association of American Medical Colleges: A nonprofit organization based in Washington, D.C. that oversees the administration of the Medical College Admission Test, hosts the American Medical College Application Service used by medical school applicants, and the Electronic Residency Application Service used by medical school students applying for residency programs. Member institutions are accredited medical schools in the United States and Canada and teaching hospitals ("About the AAMC," 2018).

At-risk Medical Student: Defined in the current study as any medical student who risks failing Step 1 of the USMLE.

Clinical Years: The portion of the medical school curriculum devoted to clerkships, commonly known as clinical rotations. Typically, two years at U.S. medical schools, but some have increased this portion of the curriculum to two and a half years, compressing the preclinical years to 18 months.

Data Mining: The use of computational techniques to identify relationships in large sets of data or predict what is likely to occur given a certain scenario (Chen & Fawcett, 2016).

Holistic Review: Refers to a medical school admission practices used to create a diverse class of students able to meet the medical needs of a growing diverse population, an important requirement to achieving health equity (Elks et al., 2018).

Machine Learning: the use of computers to provide a "good and useful approximation" of an outcome using patterns in data to make predictions (Alpaydin, 2010, p. 2).

Medical College Admission Test: A standardized, multiple-choice exam required for admission to most U.S. and Canadian medical schools ("About the MCAT exam," 2018).

Medical School: Undergraduate medical training at schools designated as allopathic, which grades the Medical Doctor degree.

National Board of Medical Examiners: An independent non-profit organization which oversees the administration of assessments for healthcare professionals ("About NBME," 2018)

Preclinical Years: The initial portion of the medical school curriculum devoted to basic sciences courses. Typically, two years at U.S. medical schools, but some have compressed this portion of the curriculum to 18 months.

Predictive Modeling: The use of data mining to create models to predict likely outcomes. When used in an educational setting predicted outcomes inform interventions designed to improve outcomes (Clow, 2013).

Step 1: The first examination of the USMLE used to evaluate the application of basic science to the practice of medicine.

Supervised Learning: to predict the value of an outcome measure based on learning from input, or predictor, variables (James, et al., 2015).

Undergraduate Grade Point Average: The cumulative grade point average from courses taken in preparation for medical school, sometimes reported in total or the undergraduate grade point average of science courses.

Underrepresented in Medicine: Medical students who classify themselves as Black, Mexican-American, Native American, or from mainland Puerto Rico ("Underrepresented," 2018).

United States Medical Licensing Exam: A standardized exam consisting of three parts, or steps, taken at different times during medical training. The first two steps are required for graduation by many medical schools in the United States. The first part, Step 1, is an examination used to evaluate the application of basic science to the practice of medicine. Step 2 assesses clinical knowledge and skills. Step 3 is the final step, required for unsupervised practice of medicine (USMLE, 2018).

List of Acronyms

AAMC: Association of American Medical Colleges

MCAT: Medical College Admission Test

NBME: National Board of Medical Examiners

ROS: Random Over Sampling

RUS: Random Under Sampling

SMOTE: Synthetic Minority Oversampling Technique

UGPA: Undergraduate Grade Point Average

URM: Underrepresented in Medicine

USMLE: United States Medical Licensing Exam

Chapter Summary

This first chapter presented the background, problem statement, research goals, and research questions. The next chapter focuses on the relevant literature in the study of Step 1 outcomes and describes the variables of interest in the current study. The six-phase process model employed to complete this study is reviewed in the third and fourth chapters. Finally, a chapter which summarizes findings from the current study and areas identified for future research.

Chapter 2

Review of the Literature

The purpose of this chapter is to demonstrate the continued application of analytics in education, referencing its origin and its place in the research community. Additionally, factors which have been previously associated with USMLE Step 1 outcomes in prior research will be reviewed as support for the predictor variables used in the current study. Finally, there will be a review of prior investigations of the factors associated with Step 1 outcomes, highlighting past methodology issues that contribute to the approach used in this study.

Analytics in Education

The use of analytic strategies in medical education allow medical school administrators to investigate student outcomes from three different time perspectives: (1) past performance to compare actual and expected student outcomes, (2) current performance to alert students as to what actions they should take to improve outcomes, and (3) predicting how students are likely to perform in the future (Ellaway, Pusic, Galbraith, & Cameron, 2014). Early references to analytics in education considered learning analytics (LA) as a mechanism by which to use the vast amount of data produced by students, for students, to "assess academic progress, predict future performance, and spot potential issues" (Johnson, Smith, Willis, Levine, & Haywood, 2011, p. 28). At this time, LA was in the early definition and application stages, yet there was a promise to harness the capabilities of data mining and modeling, concepts already used in business to uncover fraud and predict the customers who are at-risk of leaving

one company for another, as examples. In subsequent years, LA continued to evolve, with slight variations in definition. In 2012, LA was positioned as a way for teachers, in near real time, to adapt the learning process according to student need (Johnson, Adams, & Cummins, 2012). In 2013, LA was considered to be an emergent field of research, using student data to improve the learning process, and using predictive modeling algorithms to define and target at-risk populations with new retention strategies where others may have failed (Johnson, Adams Becker, Cummins, Estrada, Freeman, & Ludgate, 2013). LA was later called the educational application of big data (Johnson, Adams, Becker, Estrada, & Freeman, 2014), using data mining tools for early recognition of challenges, improve learning outcomes, and personalize learning for students as needed (Johnson, Adams Becker, Estrada, & Freemen, 2015).

As noted in the 2013 Horizon Report, LA was considered to be an emerging research discipline (Johnson et al., 2013), with big data in education as the catalyst. Early uses were focused on academic analytics to improve organizational processes and effectiveness by the adoption of business intelligence tools. Although not technically learning analytics, because of the lack of focus on student success, this was the beginning. According to Siemens (2013), the LA discipline had evolved into a collection of tools (commercially available statistical analysis products), techniques (the algorithms used in learning analytics for data mining, machine learning, and artificial intelligence), and applications (the way techniques are utilized).

Two research communities emerged from the LA discipline: Educational Data Mining (EDM), and Learning Analytics and Knowledge (LAK). The International Educational Data Mining Society focuses on the use of EDM and methods to explore data

from educational settings, and application of these methods to better understand students (Educational Data Mining, n.d.). The society supports EDM research with the annual EDM conference series and publishes the Journal of Educational Data Mining, a peer reviewed and open-access resource to address the challenges unique to EDM (Journal of Educational Data Mining, n.d.). Members of the Society for Learning Analytics Research are international researchers investigating the role of LA on teaching and learning (About SoLAR, n.d.). This society supports the LA research community by sponsoring annual LA conferences and the Learning Analytics Summer Institute. The society also publishes the Journal of Learning Analytics, a peer-reviewed and open access journal focused on the analytic challenges aimed to improve learning (Journal of Learning Analytics, n.d.). Both communities aim to improve the analysis quality of education big data, and support research and practice (Siemens & Baker, 2012). One of the main differences between EDM and LA are the techniques and methods employed by each. EDM is focused on data mining, classification, prediction, and visualization. LA uses "social network analysis, sentiment analysis, influence analytics, discourse analytics, learner success prediction, concept analysis, and sensemaking models" (Siemens & Baker, 2012, p. 253).

Analytics in education settings have remained in subsequent Horizon Reports, with the most recent edition reporting the need for machine learning to predict at-risk students and offer intervention programs designed to improve outcomes (Becker, Brown, Dahlstrom, Davis, DePaul, Diaz, & Pomerantz, 2018). The current study builds on this need by using machine learning to identify at-risk medical students, using tools, techniques, and applications, the three LA dimensions previously identified by Siemens

(2013), with software recommended by Slater, Joksimović, Kovanovic, Baker, and Gasevic (2017). Additional details about the software used in this study can be found in the next chapter.

Factors Associated with Step 1 Outcomes

Standardized examinations are often used to indicate how well students have prepared for the next step in their education and are often studied by researchers to determine how well examinations predict future academic success. For example, Scholastic Aptitude Test (SAT) scores combined with high school grade point averages indicate how well an applicant is prepared for undergraduate studies. College admission boards consider both factors when making admission decisions and have been found to be strong predictors of grades during the first year of college (Shaw, 2015). Similarly, scores from either the Graduate Record Examination (GRE) or the Graduate Management Admissions Test (GMAT) combined with UGPA have been used by graduate school admissions committees as indicators of future course grades and graduate school grade point averages (Klieger, Cline, Holtzman, Minsky, & Lorenz, 2014; Dakduk, Malavé, Torres, Montesinos, & Michelena, 2016; Klieger, Bridgeman, Tannenbaum, Cline, & Olivera-Aguilar, 2018).

In medical education, standardized examinations are used throughout the education continuum beginning with the MCAT, used to gain access to medical education, and all steps of the USMLE required to graduate from medical school, complete residency training, and become fully licensed to practice medicine. Much of the research literature in this area has been situated as predictive validity studies, investigating the degree a standardized exam score can predict future outcomes such as

preclinical course grades or USMLE step outcomes. Step 1 of the USMLE is often thought to be one of the most important milestones along the continuum because failure could lead to dismissal from medical school (Gullo et al., 2015) or elimination from the residency matching program (National Resident Matching Program, 2016). Step 1 has also been the focus of many predictive validity studies, investigating the relationship of Step 1 outcomes to future medical student clinical performance (see Burish, Fredericks, Engstrom, Tateo, & Josephson, 2015; Nagasawa et al., 2017 as examples) and the factors which best predict Step 1 outcomes, the focus of the present study. Next will be a review of the factors that have been associated with Step 1 outcomes in prior research as support for the predictor variables used in this study, position the current study in the education research literature, and to identify past methodology issues that contribute to the approach of the current study.

MCAT and UGPA are commonly used together by admissions committees to review applicants; however, each indicates cognitive ability from a different perspective (Julian, 2005). The MCAT is a point in time assessment of academic preparation for medical school. Applicants can prepare for, and may take the MCAT multiple times, and be ranked according to scores. UGPA is more school specific and a "longitudinal, continuous measure that may reflect other desired attributes, such as persistence, stamina, determination, conscientiousness, and so on" (Stratton & Elam, 2014, p. 5). As such, applicants cannot be compared according to UGPA due to differences in undergraduate school selectivity or premedical curricula but considered in tandem with MCAT scores as a "leveling factor" (Gauer et al., 2016, p. 1) and part of the holistic review process to assess the contribution of cognitive ability to the qualities needed to become a capable

physician (Capers et al., 2018). The use of MCAT and UGPA in admission decisions notwithstanding, researchers have been interested in the ability of each to predict future academic success, specifically Step 1.

Early investigations into the ability of UGPA and MCAT to predict Step 1 outcomes were focused on their predictive validity (Jones & Thomae-Forgues, 1984; Wiley & Koenig, 1996; Julian, 2005). These studies produced correlation coefficients indicating the strength of this relationship using UGPA and MCAT alone, then together, to reinforce the use of MCAT scores as a sole indicator of Step 1 outcomes. Table 1 compares the coefficients from these investigations indicating a weak to moderate relationship of UGPA to Step 1. MCAT when used alone has a moderate to strong relationship to Step 1. UGPA and MCAT combined also have a moderate to strong relationship to Step 1, adding little value to MCAT alone.

Table 1

Comparison of Correlation Coefficients of MCAT Scores and UGPA to Step 1 Outcomes

Measure	Jones & Thomae-Forgues (1984)	Wiley & Koenig (1996)	Julian (2005)
UGPA	0.37	0.48	0.49
MCAT	0.63	0.72	0.72
UGPA & MCAT	0.68	0.75	0.72
Medical schools in sample	30	16	14

Predictive validity findings from Jones and Thomae-Forgues (1984) and from Wiley and Koenig (1996) show the MCAT to be a better predictor of Step 1 scores than UGPA, with a slight increase in the correlation coefficients when using MCAT and UGPA together. Julian (2005) also found the MCAT to be a better predictor of Step 1 scores than UGPA alone but did not find an improvement in the correlation coefficients when considering MCAT and UGPA together. These investigations are representative of

MCAT validity research spanning 20 years; however, recent investigations show weaker relationships to Step 1 outcomes, adding curricular variables as a way to increase the variance explanation of Step 1 outcomes (see Table 2). The current study does not challenge the predictive validity of MCAT or UGPA but will use both as predictor variables due to the role they play in the admissions process (Capers et al., 2018).

Gender, age, and race have been found to be associated with Step 1 outcomes but are not used for acceptance decisions. Generally speaking, white males under 25 score better on Step 1 (Kleshinski et al., 2009; Andriole & Jeffe, 2010; Gauer & Jackson, 2018). Kleshinski et al. (2009) found age to be inversely related to Step 1 scores (standardized beta = - 0.10, $p < 0.009$). Nontraditional students, over 25 years old, had mean scores almost 7 points lower than students 25 years old or younger. Race was positively related to Step 1 scores, with whites scoring higher than all other races in their sample combined. Black medical students have been found to score significantly lower than other races on Step 1 (Kleshinski et al., 2009). Sesate et al. (2017) found medical students classified as underrepresented in medicine (URM) by their race to have lower Step 1 scores compared to other students ($r = - 0.32$, $p < 0.01$). URM is defined as students who are Black, Mexican-American, Native American, or from mainland Puerto Rico ("Underrepresented," 2018). Andriole and Jeffe (2010) found almost half of the URM students in their study failed Step 1 on their first attempt. The present study does not challenge the relationship of gender, age, and race to Step 1 outcomes, but includes them as predictor variables.

Recent investigations into the factors associated with Step 1 outcomes report MCAT correlation coefficients lower than earlier predictive validity research concluding

it is no longer possible to identify students who risk failing medical licensure examinations at matriculation but require preclinical course outcomes to improve prediction models (Barber, Hammond, Gula, Tithecott, & Chahine, 2018). Using preclinical course outcomes to improve correlation coefficients is not unexpected because Step 1 tests the basic sciences necessary to study medicine, and also the basis of a typical medical school preclinical curriculum (Saguil et al., 2015). To be effective, predictive models using curricular variables should identify how early at-risk students can be identified in the preclinical curriculum to give students enough time for intervention programs to be effective (Winston et al., 2014).

Findings from five recent Step 1 outcome prediction studies are summarized in Table 2, indicating the improvement in correlation coefficients from using the MCAT alone or including preclinical course outcomes in prediction models. MCAT correlation coefficients for these studies are lower than previously reported; however, prior reports include outcomes from multiple medical schools (see Table 1) prior to implementing holistic review in the admissions process. While not directly comparable to prior validity studies because of holistic review, outcomes referenced in Table 2 indicate an improvement in the correlation coefficients when preclinical course outcomes were used but differ in the measurement period found to best identify at-risk students.

Table 2

Effect of Curricular Variables on Step 1 Correlation Coefficients

Author	MCAT	Curricular Variables	Measurement Period
Saguil et al. (2015)	0.34	0.73	End of year 2
Giordano et al. (2016)	0.18	0.71	End of year 1
Khalil et al. (2017)	0.44	0.70	End of year 1
Lee et al. (2017)	0.26	0.55	End of first course
Sesate et al. (2017)	0.51	0.76	End of year 1

Note. Each of the referenced studies used students from one medical school in the sample.

Using the preclinical GPA measured at the end of the second year, Saguil et al. (2015) improved the variance explained by MCAT alone from 12% to 53%. Similarly, Giordano et al. (2016) used scores from a standardized examination given prior to Step 1 to explain 50% of the variance of Step 1 scores, an improvement of 3% using the MCAT alone. Khalil et al. (2017) used a standardized exam given at the end of the first and second years to explain 49% and 66% of the variance of Step 1 scores, an improvement over 19% using the MCAT alone. Lee et al. (2017) used the first preclinical course final grade to explain 30% of the Step 1 variance, an improvement over the MCAT alone at 7%. Finally, Sesate et al. (2017) used the end of year grade point averages for the first and second year to explain 58% and 69% of Step 1 variances, and improvement over using the MCAT alone at 26% explanation of variance. The current study does not challenge the importance of preclinical course outcomes to identify students at-risk of Step 1 failure but underscores the claim of Barber et al. (2018) that prediction of Step 1 outcomes cannot be made prior to matriculation.

It is common for medical schools to integrate Step 1 practice examinations into the preclinical curriculum. The NBME Comprehensive Basic Science Examination (CBSE) is a common practice examination used for students to determine where additional study time is needed prior to taking Step 1. The CBSE has been found to be significantly correlated with Step 1 outcomes (Giordano, Hutchinson, & Peppler, 2016), and when combined with preclinical course outcomes can explain up to 81% of the variance in Step 1 scores (Khalil et al., 2017).

Both preadmission variables and preclinical course outcomes have been found to be associated with Step 1 outcomes. Admission policies relative to minimally acceptable UGPA and MCAT scores influenced early predictive validity research, as noted in Table 1, and the ability to easily identify at-risk students. However, now that applicants with lower MCAT scores are accepted into medical school, traditional measures of cognitive ability and future outcomes are no longer valid. Medical school administrators must rely on preclinical course outcomes to identify students who are struggling.

Related Student Outcome Research

Many studies in higher education have employed a variety of machine learning techniques, moving away from inferential statistics, to predict student attrition. For example, Herzog (2006) investigated the use of different machine learning techniques to identify freshman students who are unlikely to return for their sophomore year. Using randomly selected first year student outcome data he found decision trees to outperform other techniques to identify at-risk students; however, decision trees marginally outperformed the logistic regression model.

Delen (2010) also sought to identify freshman students unlikely to return after their first year by testing the accuracy of various machine learning techniques. He used the CRISP-DM data mining process model to guide the study, using cross-validation methods to independently test his models. Delen found the imbalance between students returning and students not returning to be a problem in the accuracy of the techniques used, finding his models had high accuracy rates when identifying returning students, but low accuracy rates identifying at-risk students which was the focus of his investigation. He was able to improve the accuracy of his at-risk student predictions by randomly

selecting observations of students returning until the number of returning students equaled the number of students not returning. The imbalance problem is a recurring theme for similar studies. A detailed discussion of the CRISP-DM process model and imbalanced datasets can be found in the next chapter.

Lauría, Baron, Devireddy, Sundararaju, and Jayaprakash (2012) similarly sought to improved student retention by using machine learning techniques to identify students unlikely to return. Instead of using prediction accuracy to compare machine learning techniques, they used sensitivity and specificity rates to measure the ability of the model to correctly predict returning students and those not returning using true positive and true negative rates. After resolving the imbalance problem which also existed in their dataset, they found decision trees to outperform logistic regression and a mechanism to identify at-risk students.

Thammasiri, Delen, Meesad, and Kasap (2014) confirmed the need to have balanced datasets for machine learning when they sought to predict freshman student attrition at one university. They noted that institutional data used to analyze and predict student attrition is inherently imbalanced, and prediction models created with the majority class of students returning could produce erroneous results, especially when the interest is in the students who did not return. Using an oversampling technique to replicate the observations of students not returning until the number matched the students returning, they were able to achieve higher prediction rates with the Support Vector Machine (SVM) machine learning technique, which also outperformed logistic regression methods.

A prediction model created by Hutt, Gardener, Kamentz, Duckworth, and

D'Mello (2018) was able to accurately predict four-year graduation rates for 71% of their sample using 41,359 applicants to bachelor's degree programs provided in a national dataset using the Random Forests machine learning algorithm. Their goal was to determine the factors associated with four-year graduation rates, and to determine if data mining methods might help them better understand the factors which contribute to college success. They noted that although their model had a high accuracy rate, it only included data collected before matriculation; warning that prediction models should not be used to make admissions decisions. Student attrition is not a prevalent problem in U.S. medical schools; however, determinants of student success as indicated in this study translate well to a medical school context.

Prior studies have investigated factors which lead to Step 1 failure, using statistical models based on linear or logistic regression to indicate correlates of Step 1 performance with varying degrees of success. For example, using a combination of preadmission variables, such as gender, race, age, undergraduate institution selectivity, financial need, MCAT scores, and UGPA, the ability for these measures to explain variances in Step 1 outcomes ranged from 17% to 60% (Julian, 2005; Gohara et al., 2011; Gauer et al., 2016; Giordano et al., 2016; Lee et al., 2017). Similarly, curricular variables such as individual course grades, first and second year grade point averages, and results of Step 1 practice exams were only able to explain up to 60% of the variance in Step 1 outcomes (Burns & Garrett, 2015; Sesate et al., 2017; Lee et al., 2016). Although prior research suggests factors associated with Step 1 outcomes, there is no way to identify individual student outcomes (Lee et al., 2017). Moreover, these factors do not explain why students entering medical school with higher MCAT scores fail Step 1 for no

apparent reason (Sesate et al., 2017), or why students entering with lower MCAT scores pass Step 1 (Monroe et al., 2013).

Chapter Summary

Prior research shows the importance of identifying at-risk medical students at all points along the learning continuum. The MCAT plays a role in the admissions process but cannot be used as the sole variable to identify students who might struggle during their preclinical courses or risk failing Step 1. Studies cited in this chapter show the importance of using the MCAT to initially identify students who might need assistance to succeed in medical school and how performance in preclinical courses might better signal Step 1 outcomes. Additionally, these studies suggest the factors associated with Step 1, but there is no way to take this to a student level, looking for the outliers as identified by Lee et al. (2017). Prediction models used in higher education to improve graduation rates show promise as a foundation for the framework in the current study.

Chapter 3

Methodology

Process Model

The Cross-Industry Standard Process for Data Mining (CRISP-DM) process model was used to guide the design, development, and implementation of this study. CRISP-DM was selected because the process model was designed to make data mining projects "less costly, more reliable, more repeatable, more manageable, and faster" (Wirth & Hipp, 2000, p. 30). Additionally, it has been found to be more suitable for novice researchers (Kurgan & Musilek, 2006), and because of the repeatable processes more likely to be adopted by medical school administrators who desire to replicate this study in the future. The six phases of the CRISP-DM process model are: (1) business understanding, (2) data understanding, and (3) data preparation, (4) modeling, (5) evaluation, and (6) deployment. The phases are not meant to be prescriptive; but suggested tasks in each phase can be used based on the scope of the data mining project. Table 3 provides an overview of the CRISP-DM process model and the tasks completed in each phase for this study. The first four phases form the foundation of the methodology used to complete this study, and are described in this chapter. The evaluation phase is focused on a review of the results from the modeling phase and is described in the chapter four. The deployment phase is described in chapter five as this phase is informed by conclusions drawn from the results of this study. Actual model deployment was outside the scope of this study, but presented as recommendations for future deployment.

This chapter is focused on the first four phases of the CRISP-DM process model and is organized as follows. First, there is a brief description of the study setting, and the resources used in this study. Next, there is a discussion of the tasks applicable for this study which were completed in the first four phases of the CRISP-DM process model. Finally, a chapter summary will summarize the completed phases and introduce the evaluation phase which is described in the next chapter.

Table 3

Overview of CRISP-DM Phases and Tasks

CRISP-DM Phase	Tasks
1. Business Understanding	Determine business objectives and data mining goals Risk and contingency planning Establish success criteria Plan remaining phases
2. Data Understanding	Collect initial data Explore data Verify data quality
3. Data Preparation	Establish data sampling plan Establish data inclusion and exclusion plans Derive attributes needed to complete modeling Generate final dataset
4. Modeling	Select modeling technique Generate test design Build model Assess model accuracy
5. Evaluation	Evaluate model results in terms of business objectives and data mining goals Review overall process Determine next steps
6. Deployment	Plan deployment Produce final report

Note. Adapted from Wirth and Hipp (2000) and Kurgan and Musilek (2006)

Study Setting

Initial approval for this study was granted by the Dean of the School of Medicine at Baylor College of Medicine (BCM) and the Provost. This study was completed using non-identifying student demographic and academic achievement data provided during the admissions process, final grades from each course taken in the preclinical years, and pass or fail results from preparatory and licensure examinations. A Category 1 exemption for student consent to participate in this study was granted by the Institutional Review Boards at BCM and Nova Southeastern University under 45 CFR 46.101(b). A dataset of 514 students matriculating from 2013 to 2015 was provided by the BCM Office of Information Technology. IBM SPSS Modeler version 18.1.1 was used to complete the majority of the CRISP-DM phases.

The *U.S. News & World Report* ranks the BCM School of Medicine as the top medical school in Texas, 5th in the nation in terms of medical students entering primary care, and 16th in terms of research grant funding ("Baylor makes leap," 2018). Males and females are equally represented in the 736 full time medical students. Fifteen percent of the students are classified as being underrepresented in medicine according to self-reported ethnicities. For the 2017 academic year, BCM received 7,620 applications, interviewed 829 applicants, and admitted 185 students. Admitted students had an average UGPA of 3.88 and average MCAT scores of 35 for the old MCAT version and 517 for the 2015 revision of the MCAT ("Admissions," n.d.). Medical students consistently perform above the national average for USMLE Step 1 pass rates ("Record of success," n.d.).

Twenty-five courses make up the preclinical, or foundational sciences curriculum.

BCM is one of the few U.S. medical schools with compressed preclinical courses timelines reduced from two years to 18 months. Other schools are beginning to more toward this format as it gives an opportunity for medical students to begin their clinical rotations early. Table 4 shows the BCM preclinical curriculum divided into six blocks of courses. In addition to the preclinical curriculum BCM uses the Comprehensive Basic Sciences Examination as a way for students to assess their readiness to take Step 1.

Table 4

Preclinical Courses at Baylor College of Medicine

Block	Course Name
1	Foundations Basic to Science of Medicine Patient, Physician & Society I Integrated Problem Solving I
2	Immunologic & Pathologic Basis of Disease General Pharmacology Head and Neck Anatomy
3	Nervous System
4	Infectious Disease Behavioral Sciences
5	Patient, Physician & Society II Integrated Problem Solving II Ethics
6	Cardiology Renal Respiratory Hematology/Oncology Intro to Radiology & Lab Medicine Gastroenterology Endocrinology Genitourinary & Gynecology Genetics Age Related Topics Patient, Physician & Society III Patient Safety Transition to Clinical Rotations

Phase 1: Business Understanding

The purpose of the business understanding phase is to establish a foundation of business objectives, translating the objectives to a data mining problem, then developing a plan to achieve the business requirements (Wirth & Hipp, 2000). Three tasks were completed for this phase: (1) developed business objectives and data mining goals, (2) completed a risk assessment and a mitigation plan, and (3) established success criteria for this study. Research questions stated in chapter 1 were translated to business objectives and data mining goals, used to inform the experimental design used in future phases. A linkage between research questions, business objectives, and data mining goals ensured each research question was answered and could be evaluated in terms of completeness. Three data mining goals were created: (1) use common classification data mining algorithms to determine the variables associated with Step 1 failures, (2) use preadmission variables and courses grades as to determine the first point during the preclinical curriculum which bests identifies at-risk students, and (3) use common sampling methods to determine the method which improves the ability to identify at-risk students. Table 5 shows the linkage between research questions for this study, business objectives, and data mining goals.

Table 5

Business Objectives and Data Mining Goals

Research Question	Business Objective	Data Mining Goal
What are the factors associated with Step 1 failures?	Determine the factors associated with Step 1 failures using student outcomes from the Baylor College of Medicine Student Information System.	Use common classification data mining algorithms, determine the features associated with Step 1 failures.
Can data mining algorithms be used to identify students at risk of Step 1 failure?	Create a framework for continued use at BCM and possible in other medical schools to identify students at risk of Step 1 failure.	Using preadmission variables and courses grades as features in the dataset, determine the first point during the preclinical curriculum which bests identifies at-risk students.
How does the expected imbalance between students passing Step 1 and those failing Step 1 impact the use of data mining algorithms to identify students at risk of Step 1 failure?	Determine the best approach to address the expected Step 1 outcome imbalance problem.	Using common sampling methods, determine the method which improves the ability to identify at-risk students.

A risk assessment and a plan to mitigate risk were completed in this phase. Assessing risk at the beginning of the study allowed for alterations in the experimental design as needed to implement the mitigation plan. A data mining algorithm uses prior observations to learn the variables, or features, of the dataset which best accurately predict the outcome variable. Step 1 outcomes, specifically focused on failure, was the outcome variable used in this study; however, this highlighted a problem often found in binary classification models when predicting class outcomes, which is the imbalance between positive and negative outcomes (Branco, Torgo, & Ribeiro, 2016). Many binary classification models are unable to recognize and accurately predict the minority class,

thus special consideration should be made for this imbalance in terms of algorithm performance metrics and techniques to overcome the imbalance (Abeysinghe et al., 2018). An example of the imbalance between majority and minority classes can be found in the national pass rate for first-attempt Step 1 examinees, which is currently 96% ("USMLE Performance Data," 2017). Moreover, BCM students consistently outperform national averages; therefore, using BCM student data reduces the number of Step 1 failure observations below the 4% national average, underscoring the assertion of Kleshinski et al. (2009) that prediction of Step 1 outcomes due to this imbalance between positive and negative class outcomes continues to be problematic. Because datasets used to train prediction models should contain a sufficient number of observations, and there should be a balance of observations between Step 1 pass and failures (Ilin & Krisvtsov, 2015), the imbalance inherent in Step 1 outcomes was addressed in the risk mitigation plan as this impacted the ability to accurately determine the criteria for success and is a risk for the present study. The data sampling plan described in upcoming data preparation phase specifically addresses the Step 1 imbalance and is used to mitigate the risk described here.

Establishing criteria for success determined the performance metrics collected during the modeling phase, which were critical for the evaluation phase and assessment of the data mining goals, business objectives, and research questions. Because more importance was placed on the minority class of Step 1 outcomes (Step 1 failures), and minority class observations were expected to be at most 4% of the BCM dataset, algorithm accuracy is not suitable as a performance metric used to evaluate model success (Branco et al., 2016; Wei et al., 2017). Instead, models were evaluated based on

precision, recall, and the F-measure. Precision, also called the positive predictive value, is a measure of the model ability to predict the positive condition (Branco et al., 2016), failed Step 1 outcomes. Recall, also called the true positive rate, measures the strength of the model to predict the positive condition (Chauhan, Kaur, & Sharma, 2016), failed Step 1 outcomes. The F-measure (hereafter called F1) is the harmonic mean between recall and precision and has been found to be more useful than accuracy for model evaluation, especially when there is a class imbalance (Branco et al., 2016). F1 measures the effectiveness of the model to predict Step 1 failed outcomes.

	Predicted Outcomes	
Actual Outcomes	Step 1 Failure	Step 1 Passing
Step 1 Failure	Predicted failure for students who actually failed (True Positives)	Predicted passing for students who actually failed (False Positives)
Step 1 Passing	Predicted failure for students who actually passed (False Negatives)	Predicted passing for students who actually passed (True Negatives)

Figure 1. Confusion matrix of Step 1 passing and failing outcomes
Adapted from Hastie et al. (2015) and Thammasiri et al. (2014).

Calculating precision, recall, and F1 require a confusion or contingency matrix, a 2x2 matrix often used to display model performance measures (Thammasiri et al., 2014). Figure 1 shows the confusion matrix adapted for this study and compares the actual positive and negative conditions with their predicted counterpart. In terms of the positive and negative conditions for this study, true positives (TP) are the number of actual Step 1 failed outcomes that are predicted to be failing. True negatives (TN) are the number of actual Step 1 passing outcomes that are predicted to be passing. False positives (FP) are the number of actual Step 1 failure outcomes predicted to be passing. False negatives

(FN) are the number of actual Step 1 passing outcomes predicted to be failing. Table 6 shows these and additional metrics used to determine success are calculated using TP, FP, TN, and FN observations. The upcoming evaluation phase, found in chapter four, will provide specific details on the use of these performance measures, and associated conclusions to be drawn from them.

Table 6

Model Performance Metrics

Performance Measure	Definition
Positive Condition (P)	Failed Step 1 outcomes
Negative Condition (N)	Passed Step 1 outcomes
True Positives (TP)	Number of actual failed outcomes predicted to be failed
True Negatives (TN)	Number of actual passed outcomes predicted to be passed
False Positives (FP)	Number of actual failed outcomes predicted to be passing
False Negatives (FN)	Number of actual passed outcomes predicted to be failed
Accuracy	Percentage of correctly predicted outcomes, calculated as TP + TN / TP + FN + TN + FP
Precision	Model ability to predict failed outcomes, calculated as TP / TP + FP
Recall	The strength of the model to predict failed outcomes, calculated as TP / TP + FN
F1	The harmonic mean between precision and recall that measures model effectiveness

Note. Adapted from James et al. (2015) and Thammasiri et al. (2014)

Phase 2: Data Understanding

Collecting the initial dataset from the BCM student information system was the primary task completed during the data understand phase. Additional tasks completed were: (1) a review of the dataset to better understand the elements included, (2) a description of the contents of the dataset, (3) and a data quality assessment (Wirth & Hipp, 2000). Medical students who matriculated between 2013 and 2015 were included

in the initial dataset extracted from the BCM student information system. These years were selected based on an AAMC recommendation to limit USMLE Step score comparisons to the three most recent calendar years, currently 2015 to 2017, because Step examination content changes over time. Since the Step 1 examination is usually taken at the end of the second year of medical school and at the conclusion of the preclinical curriculum, matriculation dates between 2013 and 2015 were selected.

Table 7

Fields Requested from the BCM Student Information System

Field	Description
Matriculation Date	Date the student entered medical school
Gender	Gender reported during the application process
Age at Matriculation	Age in years calculated by Matriculation Date less Birth Date
Under Represented in Medicine (URM)	Applicants with the following self-reported races and ethnicities are considered URM: Black or African American, Hispanic/Latino, American Indian or Alaskan Native, Native Hawaiian or Pacific Islander.
Current and prior MCAT total score, and scores on each of the four subtests.	Total and subtest scores from all attempts
Total undergraduate grade point average, and science grade point average	Undergraduate grade point averages reported during the application process, and grade point averages from science courses only
Preclinical final course grades	Grades from all attempts of courses in the foundational sciences curriculum
Current and prior scores and outcomes from the Comprehensive Basic Science Examination	Scores and pass/fail outcomes from all attempts of the national standardized exam required by Baylor College of Medicine before taking Step 1
Current and prior USMLE Step 1 scores and outcomes	Score and pass/fail outcome from all attempts

Table 7 lists the data elements provided in the initial data set which included non-identifying student demographic and academic achievement data provided during the

admissions process, final grades from each course taken in the preclinical years, Comprehensive Basic Science Examination (CBSE) score, Step 1 score, and a Step 1 status indicating pass or fail. The CBSE is a national examination considered to be a readiness assessment for the Step 1 examination and is required for all BCM medical students. Scores from all attempts of the MCAT, CBSE, and USMLE Step 1 were included in the dataset.

A new version of the MCAT was offered beginning in 2015; however, BCM accepted MCAT scores from the prior version. Some of the 2015 matriculants also took the revised MCAT hoping to improve their scores. The BCM admissions committee considers scores from all MCAT attempts, using the last attempt to make acceptance decisions. For this study, only the first attempt was considered as the predictive ability of the MCAT decreases with multiple attempts (Dunleavy et al., 2013). All students matriculating in 2015 who subsequently took the Step 1 exam initially took the prior version of the MCAT. The MCAT total score and scores from each of the three MCAT subtests were also included in the initial dataset. Two of the subtests are science-based, one focused on biological sciences (BS), the other on physical sciences (PS). The last subtest is verbal reasoning (VR) requiring rapid comprehension and application of topics new to examinees.

A data quality analysis completed in SPSS Modeler revealed the following issues. Out of a total of 548 students matriculating between 2013 and 2015, 6% of the students did not have MCAT scores, 6% did not have Step 1 scores, and roughly 2% of the students did not have final course grades in all 23 preclinical courses. Missing MCAT scores are attributed to student entering the medical school as part of an early acceptance

program, with the MCAT being optional. Missing course and Step 1 scores are likely from students who left BCM at some point during the preclinical portion of their medical training due to personal or academic reasons. SPSS Modeler has a mechanism which can impute missing values in the dataset; however, since the missing values represented a small percentage of the overall dataset, the observations were deleted.

One of the noted deficiencies in prior research is the lower percentage of medical students who fail Step 1, which was observed in the BCM student data and identified as a risk in the business understanding phase. All BCM students matriculating in 2013 and 2014 passed Step 1 on the first attempt. Only two students matriculating in 2015 failed Step 1 on the first attempt. Using the recommendation made by Hu et al. (2016), students who passed Step 1 with a score within one standard deviation of the passing score were considered near failure as a way to increase the sample used for Step 1 outcome research. For this study the near failure students were considered failure for a new derived binary categorical variable representing Step 1 outcomes. Methods used to overcome this deficiency will be describe in the next phase and included in the final dataset.

Distribution of Step 1 outcomes across key variables such as UGPA and MCAT score validated the impact of holistic review for medical students matriculating between years 2013 and 2015 at BCM. Figure 2 shows the distribution of Step 1 outcomes across UGPA, which confirms the claim of Monroe et al. (2013) that students entering medical school with higher UGPA are just as likely to fail Step 1 as students with lower UGPA. Out of the 100 students with UGPA between 3.9 and 4.0, two students have Step 1 scores in the adjusted failure range within one standard deviation of the passing score. There are

also three students with UGPA below 3.4, clearly at the bottom of the UGPA range, yet they passed Step 1.

The distribution of Step 1 outcomes across MCAT scores is show in Figure 3. MCAT scores for BCM students matriculating between 2013 and 2013 are skewed toward the higher end of the score range; however, there are students with Step 1 scores in the adjust failure range who might have been predicted to pass based on MCAT scores alone. This also confirms the holistic review claim by Monroe et al. (2013) that student outcomes cannot be predicted by MCAT score alone as students with higher scores are just as likely to fail Step 1 as those with lower scores. Additionally, there are three students with MCAT scores below 25 who passed Step 1.

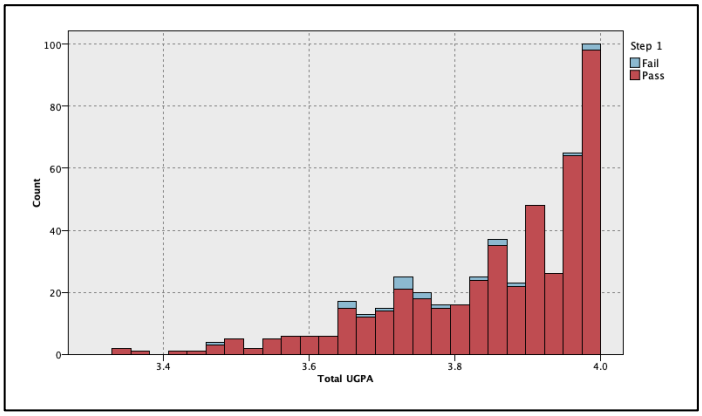

Figure 2. Distribution of Step 1 outcomes across UGPA at BCM for 2013-2015 matriculation years.

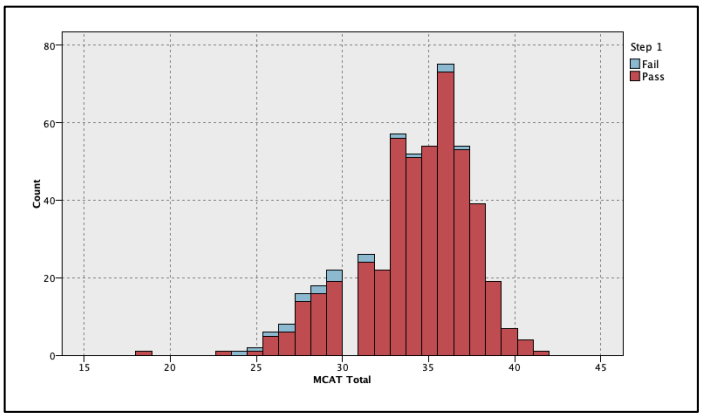

Figure 3. Distribution of Step 1 outcomes across MCAT scores at BCM for 2013-2015 matriculation years.

Phase 3: Data Preparation

The purpose of the data preparation phase is to generate final datasets used in the modeling phase, plus all associated tasks necessary to complete this phase (Wirth & Hipp, 2000). This phase consumed approximately 50% of the time needed to complete the study, as predicted by Kurgan and Musilek (2006). Four tasks were completed during this phase: (1) addressed data quality issues in the initial dataset from the BCM student information system, (2) created the new Step 1 derived outcome variable, (3) created the data sampling plan, and (4) generated the datasets to be used in the modeling phase.

The plan to exclude rows in the dataset reviewed during the prior phase is a result of the data quality audit performed in that phase. Before final datasets were created, all rows missing MCAT or Step 1 scores were removed. Additionally, rows with students who did not complete all courses were excluded from the final dataset. Due to the imbalance between Step 1 passing and failure observations it was necessary to balance the dataset so the percentage of students who failed Step 1 roughly equals the number of students who passed Step 1. The data sampling plan described next is a result of the risk assessment and mitigation plan completed in a prior phase.

One important task completed during this phase was completion of the data sampling plan. For many data mining studies, the sampling plan requires a plan for creating training, testing, and validation datasets. K-fold cross validation is a common method used to create these datasets by randomly selecting observations from the larger dataset to form *k* datasets, or folds, of equal size (Arlot & Celisse, 2010). This study used a modification of K-fold using the year of matriculation as the field to separate the folds. This modification was implemented to ensure observations from the minority class (Step 1 failures) were included in each of the three datasets. Additionally, when implemented in a medical school once the models are trained and tested, new datasets will be processed through the models by year of matriculation. For these reasons, three folds were used for this study: matriculation year 2013 for training, 2014 for testing, and 2015 for validation.

Special consideration for the imbalance between majority and minority groups of Step 1 outcomes was made based on the recommendation of Abeysinghe and colleagues (2018) and incorporated in the sampling plan. Three options were considered when developing the data sampling plan: random under-sampling (RUS), random over-sampling (ROS), and synthetic minority over-sampling technique (SMOTE). For RUS, random observances are removed from the majority class until the number of observations in both classes are approximately equal. The application of RUS in this study required random observations of passing Step 1 outcomes removed until the number of observations of passing outcomes equals the number of failing outcomes. In ROS, random observances in the minority class are selected and duplicated until both classes are about equal. As an example of application of ROS in this study, in a dataset

with 100 observations of passing Step 1 outcomes and 10 observations of failing Step 1 outcomes, a ROS plan would randomly select and duplicate of the failing rows until the number of failing observations equals 100. SMOTE requires the addition of synthetic minority class observations which are similar to other minority observations, but not exact duplicates, until both classes are approximately equal in observations (Chawla, Bowyer, Hall, & Kegelmeyer, 2002; Branco et al., 2016). This ensures observations with failed Step 1 outcomes are built with synthetic data which closely represents real data, rather than exact duplicates.

The second task completed was the creation of the derived Step 1 outcome variable based on a recalculated passing score. This was done as a way to increase the number of Step 1 failure observations based on the suggestions of Hu et al. (2016). Hu and colleagues suggested considering students who passed Step 1, but within one standard deviation of the passing score could be considered near failure for predictive modeling purposes. Table 8 indicates how students were classified according to Step 1 scores by matriculating year. Adding one standard deviation to the passing score of 192 provided a new derived passing score. Students with actual Step 1 scores below the derived passing score were recategorized with failing outcomes. This calculation was performed in IBM SPSS Modeler before the final dataset was created. Table 9 shows the effect of the new derived Step 1 outcome on the number of failed observations, increasing the number of failed observations from 2 to 20, representing a 96% passing rate which is consistent with the national average.

Table 8

USMLE Step 1 Passing Scores, Mean, and Standard Deviation by Year

Matriculation Year	Step 1 Year	Passing Score	Mean	Standard Deviation	Actual Failed Scores	Derived Failed Scores
2013	2015	192	229	20	< 192	< 213
2014	2016	192	228	21	< 192	< 214
2015	2017	192	229	20	< 192	< 213

Note. Adapted from USMLE Score Interpretation Guidelines, retrieved from *http://www.usmle.org/pdfs/transcripts/USMLE_Step_Examination_Score_Interpretation_Guidelines.pdf*.

Table 9

Effects of the Adjusted Step 1 Outcomes by Matriculating Year

Step 1 Outcome	2013	2014	2015	Total
Original Step 1 Outcome				
Step 1 Pass	180	178	156	514
Step 1 Fail	0	0	2	2
Pass Rate	100%	100%	99%	100%
Adjusted Step 1 Outcome				
Step 1 Pass	173	173	150	496
Step 1 Fail	7	5	8	20
Pass Rate	96%	97%	95%	96%

Note. Step 1 outcomes were adjusted based on recommendations by Hu et al. (2016).

To conclude this phase, six datasets were created based on the data sampling plan and the 3-fold validation plan based on year of matriculation. Four datasets were created for model training using students matriculating in 2013; one unbalanced dataset and three balanced datasets using RUS, ROS, and SMOTE. Year 2014 was used as the testing dataset and year 2015 for validation. Only the training datasets were balanced so the model can be trained first, then tested and validated with previously unseen data. As shown in Table 10, the all datasets contained preadmission variables, scores from the first

attempt of the MCAT, final course grades from all six blocks of the preclinical curriculum, score from the first attempt of the CBSE, and the derived Step 1 outcome variable.

Table 10

Variables Included in the Final Dataset

Field	Description
Matriculation Year	The year the student entered medical school, used to partition data into training, testing, and validation dataset. Not included as a potential predictor variable for Step 1 outcomes.
Gender	Gender reported during the application process
Age at Matriculation	Age in years calculated by Matriculation Date less Birth Date
Under Represented in Medicine (URM)	Applicants with the following self-reported ethnicities are considered URM: Black or African American, Hispanic/Latino, American Indian or Alaskan Native, Native Hawaiian or Pacific Islander.
MCAT total score, and scores on each of the three subtests.	Total and subtest scores from the first attempt of the MCAT. Pre-2015 scores were used for this study.
Total undergraduate grade point average, and science grade point average	Undergraduate grade point averages reported during the application process, and grade point averages from science courses only
Preclinical final course grades	Grades from the first attempts of courses in the preclinical sciences curriculum
Comprehensive Basic Science Examination Outcome	Scores and pass/fail outcomes from the first attempt of the national standardized exam required by Baylor College of Medicine before taking Step 1
Derived USMLE Step 1 outcomes	Outcome derived from adjusting the passing score down one standard deviation.

Phase 4: Modeling

The purpose of the modeling phase is to select and apply data mining algorithms, calibrating parameters within each of the algorithms to optimal values (Wirth & Hipp, 2000). Two tasks were completed during this phase: (1) generate the test design, and (2) build models according to the design. All modeling tasks were performed using IBM

SPSS Modeler using default parameters throughout. Although Wirth and Hipp recommend fine-tuning parameters to their optimal values, default parameters were used enable future replication of this study, or for ease of adoption in medical schools.

The Classification and Regression Tree (CART) classification algorithm was selected as the modeling method for this study. CART is a type of classification tree used to predict a qualitative response, such as *passing* and *failing*, when using continuous data for predictor variables (Chen et al., 2017). IBM SPSS Modeler has the ability to apply several classification algorithms, deciding which algorithm is best in terms of accuracy. This feature was not used because accuracy is not a suitable performance measure when using imbalanced datasets. Since this study is not focused on a comparison of several classification modeling techniques, only one algorithm was applied to each of the experiments, described next.

Eight experiments were included in the test design, as shown in Table 11. Each experiment used the dataset created during the prior phase, using only predictor variables specified, so final grades from the six preclinical course blocks and the final CBSE outcome could be added in the order taken by BCM students. A list of the most important variables used to create the prediction models, as determined by SPSS Modeler, will be created at the end of each experiment; however, the variables used in subsequent experiments will not be reduced to only include these variables. Important variables identified in each experiment will be used to determine the factors related to Step 1 outcomes, one of the research questions for the current study. Model performance metrics for each of the experiments are evaluated in the next phase to identify the point in the preclinical curriculum which best predicts Step 1 outcomes so appropriate

interventions can be used in an attempt to change the outcomes.

Table 11

Modeling Phase Experimental Design

Experiment	Predictor Variables
1	Preadmissions Variables (9 Predictor Variables)
2	Experiment 1 + 1st Block Course Grades (12 Predictor Variables)
3	Experiment 2 + 2nd Block Course Grades (15 Predictor Variables)
4	Experiment 3 + 3rd Block Course Grades (16 Predictor Variables)
5	Experiment 4 + 4th Block Course Grades (18 Predictor Variables)
6	Experiment 5 + 5th Block Course Grades (21 Predictor Variables)
7	Experiment 6 + 6th Block Course Grades (34 Predictor Variables)
8	Experiment 7 + CBSE Outcome (35 Predictor Variables)

Note. Each experiment used matriculation year 2013 as the training dataset, 2014 as the testing dataset, and 2015 as the validation dataset.

In addition to the stepwise approach for the use of variables, each experiment implemented the data sampling plan from the prior phase. Models were trained according to the dataset was split according to the year of matriculation, allowing for cross-validation of all models. For all experiments, matriculation year 2013 was used as the training dataset, year 2014 used as the testing dataset, and year 2015 was used as the validation dataset. Testing and validation datasets are used to present the training model observations it has not previously seen to test the prediction accuracy of Step 1 outcomes. As shown in Table 9, there are Step 1 failed outcomes for each matriculation year after adjusting the passing score down one standard deviation. RUS, ROS, and SMOTE was applied to the training dataset only.

Modeling using IBM SPSS Modeler generally followed these steps. The dataset produced in the prior phase was used for input, specifying only the fields needed during

the experiment. Three datasets were used for training, testing, and validation of each model, split by year of matriculation. Using 2014 and 2015 matriculating years allowed for testing and validation of trained models with unseen data, reducing the chances of overfitting each of the models. Performance metrics were documented for each of the experiments, used during the upcoming evaluation phase.

Chapter Summary

This chapter provided details about the methods used to complete the eight experiments described in phases 1 through 4 of the CRISP-DM process model. Each phase contributed to the to the methodology used to complete this study. Data mining goals were established, informed by the research questions for this study. A data quality audit confirmed the imbalance of classes of Step 1 outcomes. The data sampling plan provided methods to address the imbalance between Step 1 passing and failing outcomes. Models were trained, tested, and validated to follow medical student progress during the preclinical years of training. Performance metrics were gathered for each of the experiments. Evaluation of the performance metrics as specified in the fifth phase of the process model is presented in the next chapter.

Chapter 4

Results

The previous chapter concluded with the completion of the modeling phase of the CRISP-DM process model, whereby eight experiments were conducted to predict Step 1 outcomes starting with a dataset of preadmission variables, then adding preclinical course final grades for each course in the six blocks that make up the BCM preclinical curriculum, followed by results of the CBSE. This chapter focuses on evaluation of the models as specified in the fifth phase of the CRISP-DM process model and is organized as follows. Findings from each of the eight experiments are presented, comprised of the performance metrics for each of the eight experiments, a review of the outcomes for each data balancing plan, and a review of the fields identified in each experiment which best contribute to predicting Step 1 outcomes according to relative importance as determined by SPSS Modeler.

All experiments followed this general approach. A model was created for each experiment to represent the predictor variables available for students at matriculation and throughout the preclinical curriculum. Each model was trained with the CART algorithm using the original unbalanced dataset for the 2013 matriculating year (n=170, 6 failed Step 1 observations). The training dataset was then balanced using RUS (n=12, 50% failed Step 1 observations), ROS (n=328, 50% failed Step 1 observations), and SMOTE (n=328, 50% failed Step 1 observations) methods. The model was then tested with a dataset containing medical students matriculating in 2014 (n=169, 5 failed Step 1 observations) and validated with a dataset of medical students matriculating in 2015

(n=146, 8 failed Step 1 observations), both in their original unbalanced state. Full details of the methodology used to create the datasets used for each experiment and the modeling approach were discussed in Chapter 3.

Phase 5: Evaluation

The purpose of this phase is to evaluate modeling results in terms of the success criteria specified in the first phase. Each experiment was evaluated as follows. Accuracy, precision, recall, and F1 performance metrics were calculated from contingency matrices for training, testing, and validation datasets, using only the validation results for model comparison. Models were ranked according to F1, noting precision and recall for each model, which are used for further evaluation when models have identical or close F1. For example, two models with F1 of 0.40 will be evaluated by precision first indicating the ability of the model to predict Step 1 failures, then recall to evaluate the strength of the model.

Finally, top-ranking models from each experiment were reviewed to evaluate performance metric trends, using the same evaluation hierarchy for testing, training, and validation of models. Because each experiment adds new predictor variables according to preclinical course progression for BCM students, the trend will identify the point in time in which predictor variables best signal Step 1 outcomes. For example, if F1 peaks at 0.50 for experiment 4, then declines for remaining experiments, this suggests medical students at risk of Step 1 failure can best be identified by important predictor variables determined for this experiment.

Baseline Results

As shown in Table 12, baseline results were first calculated to determine

performance metrics under three scenarios: (1) all predictions match actual outcomes in the validation dataset, (2) all students were predicted to pass Step 1, and (3) all students were predicted to fail Step 1. A model in which all predictions match actual outcomes will have 100% accuracy, and 1.0 precision and recall. If all observations were predicted to pass, the resulting model would have 94.5% accuracy, but precision, recall, and F1 of 0.0. Similarly, if all observations were predicted to fail, the resulting model would have 5.5% accuracy, precision of 1.0, recall of 0.05, and F1 of 0.10. These results are unlikely but provide more of a what-if scenario in which to compare experiment results, and further illustrate why accuracy, precision, and recall alone cannot provide a full explanation of how well a model performs; however, it is expected that models will perform better than the all failure scenario in terms of F1. Results of each experiment will be compared to baseline and to results of the other experiments.

Table 12

Baseline Model Performance Metrics

Step 1 Outcomes	TP	TN	FP	FN	Accuracy	Precision	Recall	F1
Actual Outcomes	8	138	0	0	100.0%	1.00	1.00	1.00
All Passing	0	138	8	0	94.5%	0.00	0.00	0.00
All Failures	8	0	0	138	5.5%	1.00	0.05	0.10

Experiment 1

The purpose of this experiment was to create a model to predict Step 1 outcomes using preadmission variables only, representing variables available to medical school administrators at matriculation. Table 12 summarizes model performance metrics for the validation dataset using models trained with unbalanced and balanced datasets. Testing, training, and validation performance metrics for this experiment can be found in

appendix A.

The model trained with the unbalanced dataset was not able to correctly predict any of the failed observations in the validation dataset yet was able to accurately predict 94.5% of the passing outcomes. The model trained using RUS was able to accurately predict six of the eight failed observations, with 0.75 precision, 0.10 recall, and F1 of 0.17. The model trained using ROS was able to accurately predict two of the eight Step 1 failing observations, with 0.25 precision, 0.20 recall, and F1 of 0.22. Finally, the model trained using SMOTE was able to accurately predict two of the eight failed observations, with 0.25 precision, 0.15 recall, and F1 of 0.19. Using F1 to evaluate the effectiveness of each model, the ROS training dataset slightly outperformed the other balancing methods; but the RUS model accurately predicted the most failed Step 1 observations.

Table 13

Experiment 1 Model Performance Metrics by Balance Method

Method	TP	TN	FP	FN	Accuracy	Precision	Recall	F1
Unbalanced	0	138	8	0	94.5%	0.00	0.00	0.00
RUS	6	81	2	57	59.6%	0.75	0.10	0.17
ROS	2	130	6	8	90.4%	0.25	0.20	0.22
SMOTE	2	127	6	11	88.4%	0.25	0.15	0.19

Note. Results using 2015 matriculating year as the validation dataset (n=146, 8 failed Step 1 observations). RUS = Random Under Sampling, ROS = Random Over Sampling, SMOTE = Synthetic Minority Oversampling Technique, F1 = the harmonic mean of precision and recall representing model effectiveness.

SPSS Modeler reflects predictor importance as a relative measure compared to the other predictor variables, not a correlation coefficient to measure the strength of the relationship between the predictor variable and Step 1 outcomes. Of the 9 predictor variables used in this experiment, the RUS balanced dataset determined UGPA as the most important with a relative importance of 0.69. The decision tree for the RUS dataset

indicated students with a UGPA less 3.886 would fail Step 1. The ROS balanced dataset determined VR and UGPA to be the most important predictor variables with a relative importance of 0.49 and 0.30 respectively. VR as the most important predictor variable was not expected as this portion of the MCAT does not test for biological or physical science knowledge needed to pass Step 1. The ROS decision tree indicates VR scores greater than 8.5 and UGPA greater than 3.725 will pass Step 1. The SMOTE balanced dataset determined UGPA and VR to be equally important with a relative importance factor of 0.37. The SMOTE decision tree indicates VR scores greater than 8.998 and UGPA greater than 3.787 will pass Step 1.

Based on F1, the ROS model outperforms the other balancing methods, and outperforms the baseline expectation of 0.10 F1. Further comparisons and conclusions will be made as predictor variables are added. Findings for the next experiments will be presented similar to this but will compare the performance metrics for the most effective model (as determined by F1) to conclude the point in the preclinical curriculum which best predicts medical students at risk of failing Step 1. Subsequent experiments will begin to add final preclinical course grades in order of the blocks of courses taken by BCM students.

Experiment 2

In addition to the preadmission variables used in the prior experiment, training, testing, and validation datasets also contain final grades for the following courses: Foundations Basic to the Science of Medicine (FBS), Patient, Physician & Society – Part 1 (PPS1), and Integrated Problem Solving – Part 1 (IPS1). Table 13 summarizes model performance metrics for the validation dataset using models trained with unbalanced and

balanced datasets. Testing, training, and validation performance metrics for this experiment can be found in appendix B.

As found in experiment 1, the model trained with the unbalanced dataset was not able to correctly predict any of the failed observations in the validation dataset, yet was able to accurately predict 94.5% of the passing outcomes. The model trained using RUS was able to accurately predict six of the eight failed observations, with 0.75 precision, 0.10 recall, and F1 of 0.17. The model trained using ROS was able to accurately predict three of the eight Step 1 failed observations, with 0.38 precision, 0.75 recall, and F1 of 0.50. Finally, the model trained using SMOTE was able to accurately predict four of the eight failed observations, with 0.50 precision, 0.67 recall, and F1 of 0.57.

Table 14

Experiment 2 Modeling Results by Balance Method

Method	TP	TN	FP	FN	Accuracy	Precision	Recall	F1
Unbalanced	0	138	8	0	94.5%	0.00	0.00	0.00
RUS	6	81	2	57	59.6%	0.75	0.10	0.17
ROS	3	137	5	1	95.9%	0.38	0.75	0.50
SMOTE	4	136	4	2	95.9%	0.50	0.67	0.57

Note. Results using 2015 matriculating year as the validation dataset (n=146). RUS = Random Under Sampling, ROS = Random Over Sampling, SMOTE = Synthetic Minority Oversampling Technique, F=the harmonic mean of precision and recall representing model effectiveness.

Of the twelve predictor variables used in this experiment, the RUS balanced dataset determined UGPA as the most important with a relative importance of 0.69. The decision tree for the RUS dataset indicated students with a UGPA less 3.885 would fail Step 1. The ROS balanced dataset determined the FBS final grade, VR, and UGPA to be the most important predictor variables with a relative importance of 0.37, 0.36, and 0.14 respectively. FBS as the most important predictor variable was expected based on prior

research finding an improvement in prediction models when curricular variables were added. As in experiment 1, VR as an important predictor variable was not expected as the knowledge required for this portion of the MCAT does not match the content tested in Step 1. The ROS decision tree indicates FBS grades above 85.350 will pass Step 1. The SMOTE balanced dataset determined FBS to be the most important with a relative importance factor of 0.415. The SMOTE decision tree indicates FBS grades greater than 85.35 will pass Step 1.

Table 15

Top Model Performance Metrics for Experiments 1 and 2

Experiment	TP	TN	FP	FN	Accuracy	Precision	Recall	F1
1: ROS	2	130	6	8	90.4%	0.25	0.20	0.22
2: SMOTE	4	136	4	2	95.9%	0.50	0.67	0.57

Note. Results in this table are based on validation models from each experiment.

Using F1 to evaluate the effectiveness of each model, the SMOTE training dataset outperformed the other balancing methods in the ability to predict both passing and failing Step 1 observations and exceeds baseline expectations. A comparison of the top performing models in terms of F1 for the first two experiments is shown in Table 15. Comparing the top performing models in experiments 1 and 2, adding course grades for the first block of preclinical courses improved the effectiveness of the model. Accuracy has also increased, but this metric is based on the correct number of passing Step 1 observations in each experiment. Based on the findings from this experiment, students at risk of failing Step 1 can be best predicted at the end of the first block of preclinical courses, using the final FBS grade as the predictor.

Experiment 3

In addition to the preadmission variables and first block of preclinical courses used in the prior experiment, training, testing, and validation datasets also contain final grades for the following courses: Immunologic and Pathologic Basis of Disease (IPD), General Pharmacology (PHR), and Head and Neck Anatomy (HNA). Table 16 summarizes model performance metrics for the validation dataset using models trained with unbalanced and balanced datasets. Testing, training, and validation performance metrics for this experiment can be found in appendix C.

Table 16

Experiment 3 Modeling Results by Balance Method

Method	TP	TN	FP	FN	Accuracy	Precision	Recall	F1
Unbalanced	0	138	8	0	94.5%	0.00	0.00	0.00
RUS	6	81	2	57	59.6%	0.75	0.10	0.17
ROS	3	136	5	2	95.2%	0.38	0.60	0.46
SMOTE	8	123	0	15	89.7%	1.00	0.35	0.52

Note. Results using 2015 matriculating year as the validation dataset (n=146). RUS = Random Under Sampling, ROS = Random Over Sampling, SMOTE = Synthetic Minority Oversampling Technique, F1 = the harmonic mean of precision and recall representing model effectiveness.

Similar to prior experiments, the model trained with the unbalanced dataset was not able to correctly predict any of the failed observations in the validation dataset, yet was able to accurately predict 94.5% of the passing outcomes. The model trained using RUS was able to accurately predict six of the eight failed observations, with 0.75 precision, 0.10 recall, and F1 of 0.17, which is not an improvement from experiment 2. The model trained using ROS was able to accurately predict 3 of the eight Step 1 failed observations, with 0.38 precision, 0.60 recall, and F1 of 0.46, a slight decrease from experiment 2. Finally, the model trained using SMOTE was able to accurately predict all

of the eight failed observations, with 1.00 precision, 0.35 recall, and F1 of 0.52, which is an improvement in the ability of the model to predict failed Step 1 outcomes, but a slight decrease in model strength and effectiveness.

Of the 15 predictor variables used in this experiment, the RUS balanced dataset determined UGPA as the most important with a relative importance of 0.3381. The decision tree for the RUS dataset indicated students with a UGPA less 3.885 would fail Step 1. The ROS balanced dataset determined the IPD, added for this experiment, and PP1, new for this experiment, final grade to be the most important predictor variables with a relative importance of 0.59 and 0.24 respectively. All preadmission variables used in the first experiment had a relative performance factor less than 0.05 for this experiment. The ROS decision tree indicates students with a final grade less than 79.068 will fail Step 1 unless they have a PP1 final course grade over 95.25. The SMOTE balanced dataset determined IPD to be the most important predictor variable with a relative importance factor of 0.93. The SMOTE decision tree indicates students with IPD final grades greater than 83.45 will pass Step 1.

Table 17

Top Model Performance Metrics for Experiments 1 – 3

Experiment	TP	TN	FP	FN	Accuracy	Precision	Recall	F1
1: ROS	2	130	6	8	90.4%	0.25	0.20	0.22
2: SMOTE	4	136	4	2	95.9%	0.50	0.67	0.57
3: SMOTE	8	123	0	15	89.7%	1.00	0.35	0.52

Note. Results in this table are based on validation models from each experiment.

All three oversampling techniques beat baseline expectations; however, the model trained with the SMOTE dataset outperformed the other methods in terms of F1. As shown in Table 17, comparing the top performing models in the first 3 experiments

adding course grades for the first block of preclinical courses improved the effectiveness of the model, but slightly declined when the second block was added. However, with the second block of preclinical courses added as predictor variables, the SMOTE model was able to predict all failed Step 1 outcomes.

Experiment 4 & 5

Results for experiments 4 and 5 were identical. The Nervous System (NRS) course was added to experiment 4, and Infections Disease (IND) and Behavioral Sciences (BES) for experiment 5, building on to the input files used for prior experiments. Performance metrics for both experiments are summarized in Table 18. Testing, training, and validation performance metrics for this experiment can be found in appendices D and E, separated due to slight variations in training and testing performance metrics, and relative performance ratings on important predictors and decision trees.

Table 18

Experiments 4 - 5 Modeling Results by Balance Method

Method	TP	TN	FP	FN	Accuracy	Precision	Recall	F1
Unbalanced	0	138	8	0	94.5%	0.00	0.00	0.00
RUS	6	81	2	57	59.6%	0.75	0.10	0.17
ROS	1	132	7	6	91.1%	0.13	0.14	0.13
SMOTE	1	132	7	6	91.1%	0.13	0.14	0.13

Note. Results using 2015 matriculating year as the validation dataset (n=146). RUS = Random Under Sampling, ROS = Random Over Sampling, SMOTE = Synthetic Minority Oversampling Technique, F1 = the harmonic mean of precision and recall representing model effectiveness.

Consistent with prior experiments, the model trained with the unbalanced dataset was not able to correctly predict any of the failed observations in the validation dataset yet was able to accurately predict 94.5% of the passing outcomes. The model trained using RUS was able to accurately predict six of the eight failed observations, with 0.75

precision, 0.10 recall, and F1 of 0.17, which is not an improvement from experiment 3. The model trained using ROS was able to accurately predict 1 of the eight Step 1 failed observations, with 0.13 precision, 0.14 recall, and F1 of 0.13, a decrease from experiment 3. Finally, the model trained using SMOTE was able to accurately predict 1 of the eight failed observations, with 0.13 precision, 0.14 recall, and F1 of 0.13, also a decrease from the prior experiment.

The RUS balanced dataset determined UGPA as the most important with a relative importance of 0.3381. Preclinical courses added for these experiments did not affect predictor importance or the decision trees, as they were not selected as important variables. The decision tree for the RUS dataset indicated students with a UGPA less 3.886 would fail Step 1. The ROS balanced dataset determined the NRS final grade, added in experiment 4, to be the most important predictor variables with a relative importance of 0.94. The ROS decision tree indicates students with an NRS final grade less than 80.050 will fail Step 1. The SMOTE balanced dataset also determined NRS to be the most important predictor variable with a relative importance factor of 0.96. The SMOTE decision tree indicates students with an NRS final grade less than 80.050 will fail Step 1.

All three oversampling methods continued to beat baseline expectations, but in terms of F1 the RUS model slightly outperformed the models and had the highest number of accurate Step 1 failed observations; however, the model did not find the variables added in either experiment to contribute to Step 1 predictions. Both ROS and SMOTE found NRS to have the highest relative importance. Table 19 shows performance metrics for the first 5 experiments. Model effectiveness peaked with experiment 2 with a model

consisting of preadmission variables and final course grades for the first block of preclinical courses, dropping slightly when the second block of course grades were added, but improved precision.

Table 19

Top Model Performance Metrics for Experiments 1 – 5

Experiment	TP	TN	FP	FN	Accuracy	Precision	Recall	F1
1: ROS	2	130	6	8	90.4%	0.25	0.20	0.22
2: SMOTE	4	136	4	2	95.9%	0.50	0.67	0.57
3: SMOTE	8	123	0	15	89.7%	1.00	0.35	0.52
4: RUS	6	81	2	57	59.6%	0.75	0.10	0.17
5: RUS	6	81	2	57	59.6%	0.75	0.10	0.17

Note. Results in this table are based on validation models from each experiment.

Experiment 6, 7 & 8

Model performance metrics from experiments 6, 7, and 8 were identical and reported together. For experiment 6, final grades from the block 5 preclinical courses, Integrated Problem Solving – Part 2 (IPS2), Patient, Physician & Society – Part 2 (PPS2), and Ethics (ETH), were added as predictor variables. Final grades from the following block 5 courses were added as predictor variables for experiment 7: Cardiology (CAR), Renal (RNL), Respiratory (RSP), Hematology/Oncology (HMO), Introduction to Radiology & Lab Medicine (RLM), Gastroenterology (GST), Endocrinology (END), Genitourinary/Gynecology (GUG), Genetics (GNT), Age Related Topics (ART), Patient, Physician & Society – Part 3 (PP3), Patient Safety (PSA), and Transition to Clinical Rotations (ITC). The Comprehensive Basic Sciences Examination (CBSE) was added for experiment 8. Performance measures for all three experiments were identical and shown in Table 20, but full testing, training, and validation performance metrics for both

experiments can be found in appendices F, G, and H, separated due to differences in predictor importance.

Table 20

Experiments 6 – 8 Modeling Results by Balance Method

Method	TP	TN	FP	FN	Accuracy	Precision	Recall	F1
Unbalanced	0	138	8	0	94.5%	0.00	0.00	0.00
RUS	6	81	2	57	59.6%	0.75	0.10	0.17
ROS	1	132	7	6	91.1%	0.13	0.14	0.13
SMOTE	1	136	7	2	93.8%	0.13	0.33	0.18

Note. Results using 2015 matriculating year as the validation dataset (n=146). RUS = Random Under Sampling, ROS = Random Over Sampling, SMOTE = Synthetic Minority Oversampling Technique, F1 = the harmonic mean of precision and recall representing model effectiveness.

In each experiment the model trained with the unbalanced dataset was not able to correctly predict any of the failed observations in the validation dataset yet was able to accurately predict 94.5% of the passing outcomes. The model trained using RUS was able to accurately predict six of the eight failed observations, with 0.75 precision, 0.10 recall, and F1 of 0.17. The model trained using ROS was able to accurately predict 1 of the eight Step 1 failed observations, with 0.13 precision, 0.14 recall, and F1 of 0.13. Finally, the model trained using SMOTE was able to accurately predict 1 of the eight failed observations, with 0.13 precision, 0.33 recall, and F1 of 0.18.

Important predictor variables for this experiment are identical to experiments 4 and 5, that is the RUS balanced dataset determined UGPA as the most important with a relative importance of 0.3381. Preclinical courses added for these experiments did not affect predictor importance as they were not selected as important variables. The decision tree for the RUS dataset indicated students with a UGPA less 3.886 would fail

Step 1. The ROS balanced dataset determined the NRS final grade, from the third block of preclinical courses added in experiment 4, to be the most important predictor variables with a relative importance of 0.94. The ROS decision tree indicates students with an NRS final grade less than 80.050 will fail Step 1. The SMOTE balanced dataset also determined the third block NRS course to be the most important predictor variable with a relative importance factor of 0.96. The SMOTE decision tree indicates students with an NRS final grade less than 80.050 will fail Step 1. None of the predictor variables added for these experiments were ranked as important variables. Comparing models on F1 indicate the SMOTE model as the top performing model for experiments 6, 7, and 8.

Table 21

Top Model Performance Metrics for Experiments 1 – 8

Experiment	TP	TN	FP	FN	Accuracy	Precision	Recall	F1
1: ROS	2	130	6	8	90.4%	0.25	0.20	0.22
2: SMOTE	4	136	4	2	95.9%	0.50	0.67	0.57
3: SMOTE	8	123	0	15	89.7%	1.00	0.35	0.52
4: RUS	6	81	2	57	59.6%	0.75	0.10	0.17
5: RUS	6	81	2	57	59.6%	0.75	0.10	0.17
6: SMOTE	1	136	7	2	93.8%	0.13	0.33	0.18
7: SMOTE	1	136	7	2	93.8%	0.13	0.33	0.18
8: SMOTE	1	136	7	2	93.8%	0.13	0.33	0.18

Note. Results in this table are based on validation models from each experiment.

Chapter Summary

A comparison of the top performing models for all experiments is shown in Table 21. Model effectiveness peaked with experiment 2 using a prediction model consisting of preadmission variables and final course grades for the first block of preclinical courses, dropping slightly when the second block of course grades were added, then a sharp drop

in model effectiveness for remaining experiments. Findings from this study suggest that identifying students at risk of failing Step 1 can be predicted as early as the end of the first block of preclinical courses, with subsequent blocks of courses adding little value to prediction models. The next chapter will summarize this study and present conclusions which can be drawn from all experiments. Additionally, the last phase of the CRISP-DM process model, the deployment phase, will be discussed in terms recommendations for deployment informed by limitations identified for the current study and opportunities for additional research.

Chapter 5

Conclusions, Implications, Recommendations, and Summary

The purpose of this chapter is to explore conclusions from findings presented in the previous chapter, and to establish the implications to BCM and to other U.S. medical schools. This chapter is organized as follows. First is a brief summary of the problem investigated in this study with a summary of conclusions presented in terms of the research questions stated in the first chapter. Next is a discussion of the implications of this study to the medical student outcomes literature and implications to U.S. medical schools in admissions and promotions processes. Recommendations for future research are then discussed, including recommendations based on limitations of the present study. Finally, a summary of this study concluding with the final deployment phase of the CRISP-DM process model.

Conclusions

Conclusions drawn from findings in the previous chapter are stated in terms of the three research questions which guided the design of this study. The first question asked for the factors associated with Step 1 failures. Important variables related to these failures were identified in each of the eight experiments; however, the most effective model based on the F1 measure should be used to answer this question. The model with the highest F1 was from experiment 2 which contained preadmission variables and the first block of preclinical courses. Full results for this experiment can be found in Appendix B. The final course grade from the FBS course was determined to be the most important predictor variable with a factor of 0.4015, followed by Science UGPA with a

factor of 0.1845, scores from the verbal reasoning part of the MCAT with a factor of 0.1769, age at matriculation with a factor of 0.1360, scores from the biological and physical sciences portion of the MCAT with factors of 0.0617 and 0.0144 respectively. MCAT total score, final course grade from the PP1 course in block 1, UGPA, and URM status each had factors of 0.0063. The FBS final grade was expected and is consistent with prior research finding curricular variables to be better predictors of Step 1 outcomes than preadmission variables alone (Saguil et al., 2015; Giordano et al., 2016; Khalil et al., 2017; Lee et al., 2017; Sesate et al., 2017), as also show in Table 2 from chapter 2. FBS is also a significant milestone in the BCM preclinical curriculum as students learn the basic sciences which serve as foundations for the practice of medicine. Therefore, it is not surprising that the Science UGPA was determined to be the second important variable from this experiment.

This is an important finding in terms of predicting Step 1 outcomes at BCM as UGPA and MCAT scores were once used as indicators of medical students at risk of Step 1 failure. As shown in Figures 2 and 3, students with higher MCAT scores and UGPA are just as likely to fail Step 1. Lee et al. (2017) described the elusiveness of prediction models which identified factors related to successful Step 1 outcomes, but also provided insight into individual student performance. Through the use of data mining and predictive models, this study provided a framework to determine likely student outcomes, regardless of MCAT or UGPA at other U.S. medical schools, adjusted for each schools' preclinical courses.

The second research question asked if data mining algorithms can be used to identify medical students at risk of Step 1 failure. This study employed the CART data

mining algorithm to build models using preadmission and curricular variables as predictor variables, and Step 1 pass or fail status as the outcome variable. Predictor variables were introduced to the model in a stepwise fashion following medical student progression through the preclinical curriculum at BCM. Three metrics were used to evaluate model performance: precision, recall, and the F-measure. Precision indicates the ability of models to accurately predict failed Step 1 outcomes. Recall measures the strength of the models, and the F-measure is a harmonic mean between both indicating model effectiveness. Model effectiveness was the strongest with models using preadmission and final course grades from the first block of preclinical courses, suggesting Step 1 outcomes can be predicted as early as the first term of the first year of medical training using data mining algorithms. Therefore, yes, data mining algorithms can be used to identify students at risk of Step 1 failure.

The last research question asked if the expected imbalance between students passing Step 1 and those failing Step 1 impacted the use of data mining algorithms to identify students at risk of Step 1 failure. Because more medical students pass Step 1 than fail (current national pass rate is 96%), special consideration was made to handle the imbalance when presenting datasets to train models. Step 1 outcomes have been more difficult to predict in the past due the low number of failed observations expected at any U.S. medical school (Hu et al., 2016). This was a challenge in the current study as BCM students traditionally performed better than national average on Step 1, which reduced the number of failed observations in data received from the student information system. Using the recommendation by Hu et al. (2017), BCM students who passed Step 1 within one standard deviation of the passing score were considered failed observations for this

study. The adjusted Step 1 pass or fail status was derived during the data preparation phase of this study.

Models in each experiment were trained using the original unbalanced dataset; but none were able to predict any of the Step 1 failures, confirming the assertion by Fernandez et al. (2018) that prediction models using binary outcome variables with an imbalance between the majority and minority classes must be balanced. This study used one undersampling and two oversampling techniques for models to learn the factors associated with failing Step 1. First, undersampling was used to randomly reduce the number of passing Step 1 observations until the training dataset contained an equal number of passing and failing observation. The RUS approach is often criticized for the omission of valuable training observations, but the final result is a balanced training dataset. For this study, undersampling outperformed other balancing methods for 25% of the experiments in terms of precision and F1. Second, oversampling was used to replicate the failing observations until the number of failing observations equal the number of training observations. Critics of the ROS approach believe duplication of failing observations does not give sufficient variation in failing observations in which the model can learn, resulting in overfitting of models. For this study, this oversampling approach outperformed other balancing methods in 12.5% of the experiments. And third, another form of oversampling was used to create synthetic examples of the failing observations which closely resemble original observations, until the number of passing and failing observations are equal. The SMOTE approach is often considered the preferred sampling method when presented with imbalanced datasets (Fernández, Garcia, Herrera, & Chawla, 2018). For this study, this oversampling approach outperformed

other balancing methods in 62.5% of the experiments. In total, oversampling outperformed undersampling in 75% of the experiments.

The imbalance between passing and failing Step 1 observations did not prevent the use of data mining algorithms to predict student Step 1 outcomes. However, findings from each of the experiments show the necessity of oversampling as models trained with the original dataset were never able to identify actual Step 1 failure observations, yet these models had the highest accuracy rates in each of the experiments. This underscores the need for additional model performance measures as accuracy rates in models trained with unbalanced datasets reflect prediction accuracy for the most observations.

To summarize, USMLE Step 1 student outcomes can be predicted using data mining algorithms, but the dataset used to train the model must be balanced using an oversampling method. At BCM, at-risk medical students can be identified by the end of the FBS course, a foundational sciences course in the first term of the preclinical curriculum. The decision tree for this model indicates Step 1 passing outcomes for students with a final grade above 85.35 in the FBS course. Findings from this study are applicable to BCM only, and cannot be generalized to other medical schools due to differences in curricula and missions. However, the method utilized in this study can be used by other faculty and administrators at other medical schools using predictor variables consistent with their preclinical curriculum.

Implications

Medical school faculty want all students to have successful outcomes during their medical education. Students matriculating into medical school did so at the recommendation of admission committee members who determined that each student is

capable of withstanding the rigor of medical training and is likely to pass licensure examinations. Although desirable, it is not feasible that all students will graduate from medical school and those who graduate could encounter difficulty along the way. However, findings from this study have implications which are applicable to BCM and other U.S. medical schools, and also contributes to Step 1 outcomes research.

Many schools have implemented programs to help students achieve successful outcomes. Holistic review admissions practices have resulted in diverse medical school classes able to meet the future medical needs of a diverse population (Elks et al., 2018); however, the change in applicant review means UGPA and MCAT scores can no longer be the only factors used to identify students for these programs (Capers et al., 2018). Based on the findings of this study, BCM faculty should closely monitor course outcomes during the initial term of the first year of medical training, specifically for the Foundations for the Basics of Medicine course. Decision rules indicated students with final grades in this course below 85.36 will fail Step 1, and could benefit from programs designed to improve overall medical school grades and outcomes on licensure examinations. Similar models using students from other medical schools will need to be adapted to the specific curriculum at each school, using the stepwise model approached used in this study.

Findings from this study also have implications for future research as the methodology used resolved three problems identified from prior research. The first problem is the lack of Step 1 failure data due to the high national pass rates (Kleshinski et al., 2009). Sample size relative to Step 1 failures will continue to be small because of high pass rates, but since pass rates vary by school, researchers replicating this study in

the future can choose how they implement the near failures (passing within one standard deviation) in their models. The design of this study used a binary categorical predictor variable for Step 1 outcomes, but future investigations could expand the predictor variable to three categories: pass, nearly pass, and fail.

The second problem is the inability to translate correlates of Step 1 failure to individual student performance prediction (Lee et al., 2017). Findings from this study show how data mining can be used to identify factors related to Step 1 failure, but also provide early identification of students at risk of failure so they can take advantage of support programs designed to improve Step 1 outcomes. Future research can focus on individual student performance, addressing one of the issues raised by Monroe and colleagues (2013) relative to the inability to explain why some students with higher academic credentials fail Step 1 for no reason, and some students with lower academic credentials pass Step 1.

The third problem is the incorrect use of model accuracy that is inherent in models trained with imbalanced datasets (Thammasiri et al., 2014). Findings from this study showed how models trained with unbalanced datasets tend to have higher accuracy rates, but model accuracy is skewed because accuracy is calculated using the outcome which is not the focus of investigation. Models should be evaluated on a combination of precision, recall, and the F1 measure (harmonic mean of precision and recall) to determine the highest performing models. Prediction models evaluated on model accuracy rates alone will not adequately identify students at risk of future Step 1 failure.

Recommendations

There are barriers and limitations to this study, identified in the first chapter and

expanded here to include opportunities for future research. Holistic review and mission-driven admissions allow medical school admissions committees to find qualified applicants who are academically prepared, have demonstrated the qualities necessary to become good physicians, and have future career aspirations which meet the unique mission of the medical school (Ellaway et al., 2018). Predictor variables in this study were specific to the preclinical curriculum at BCM, therefore, findings are delimited to this participating school only. This study could be repeated at other allopathic U.S. medical schools using predictor variables modified according to the preclinical curriculum used at these schools.

This study used scores from an older version of the MCAT, which was revised in 2015. Results of analysis completed in the data preparation phase of the CRISP-DM process model revealed all BCM students matriculating between 2013 and 2015 took the old MCAT. Some applicants elected to take the new version of the MCAT in an attempt to improve their score, but scores from the first attempt only were used for this study. Some of the models in this study identified scores from the verbal reasoning part of the old MCAT as important predictor variables; however, this does not translate to the new MCAT as this section was not carried forward. This study should be repeated in 2020 when scores from the old MCAT are not accepted at any medical school and enough time has passed for these applicants to have taken Step 1.

Only final grades from preclinical courses and the CBSE were included as predictor variables in this study, as these were readily available in the BCM student information system. Findings from this study indicate at-risk students could be identified as early as the end of the first term; however, adding outcomes from low and high-stake

assessments during the first term as predictor variables might allow faculty to identify at-risk students earlier. Doing so increases the work to be done during the data preparation phase but could add meaningful and important predictor variables to the model. More research is needed to better understand how other variables, such as the use of student-initiated study groups, USMLE preparation courses, and self-directed study behaviors effect Step 1 outcomes.

Measures of student engagement were not included as predictor variables in this study, but could be included in future research. Findings from research using course attendance as measure of engagement and association with academic performance has yielded mixed results and remains a source of controversy in medical education. Attendance has been found to be an early signal of student retention issues in higher education (Gray & Perkins, 2019). In medical education, many lecturers require students to attend class; however, as more lectures are available online, students have opted to watch lectures outside of class rather than attend in person. Learning styles and preferences notwithstanding, Eisen and colleagues (2015) did not find attendance to be associated with academic performance in preclinical courses. More research is need to determine if student engagement, whether by attendance or other measures, can be used as an early warning signal of Step 1 outcomes, especially as medical schools incorporate problem-based and active learning in the medical school curriculum.

Although not a limitation to this study, Step 1 pass or fail status was the outcome variable used, applicable to U.S allopathic medical schools. Further research is needed to determine if the methods used in this study translate to other academic medical centers with different curricula, and those who require other licensure examinations, for example

the NCLEX examination for nurse licensure, the COMLEX licensure examination for osteopathic medical students, or the PANCE for physician assistant licensure.

Summary

Step 1 of the United States Medical Licensing Examination (USMLE) is part of a three-step examination to obtain full medical licensure. Taken at the end of the preclinical portion of the medical education curriculum, passing Step 1 is required for promotion to later years in medical training, is necessary to qualify for the additional step examinations, and has been cited as the most important factor for acceptance into many of the top graduate medical education programs by residency program directors (National Resident Matching Program, 2018). Identifying the factors associated with medical students who fail Step 1 of the USMLE has been a focus of investigation for many years. Some researchers believe lower scores on the Medical Colleges Admissions Test (MCAT) are the sole factor used to identify failure (Gauer et al., 2016). Other researchers believe lower course outcomes during the first two years of medical training are better indicators of failure (Sesate et al., 2017). Yet, there are medical students who fail Step 1 of the USMLE who enter medical school with high MCAT scores, and conversely medical students with lower academic credentials who are expected to have difficulty passing Step 1 but pass on the first attempt. This phenomenon has been attributed to a holistic review of applicants, which considers life experiences and demonstrated qualities necessary to become good physicians in addition to academic qualifications (Monroe et al., 2013; "Holistic Review," 2018). Today's medical students are no longer solely determined to be at-risk of poor outcomes based on academic credentials at matriculation, as students entering medical school with lower MCAT scores

and UGPA are just as likely to be successful in medical school (Sesate et al., 2017).

The goal of this study was to identify the factors related to Step 1 failure, and to predict individual student outcomes without using MCAT scores or UGPA as sole indicators. Prior investigations have attempted to predict Step 1 outcomes, but researchers have found the low sample size due to the high national pass rate of 96% to be a limiting factor in their findings (Kleshinski et al., 2009). Moreover, prior research has found factors correlated with Step 1 outcomes but has failed to provide insight into individual student performance (Hu et al., 2016; Lee et al., 2017). Predictive modeling using data mining methods was used to identify medical students at risk of Step 1 failure, relying on computational techniques to resolve the low sample size issue and to identify Step 1 outcome relationships in sets of data (Chen & Fawcett, 2016). Similar models have been used in business settings to find new customers, or to identify customers likely to stop using a product or service (Lee, Kim et al., 2017; Zhao et al., 2017). Predictive models have also been applied in education settings to predict high school dropouts and improve university student retention (Thammasiri et al., 2014; Marquez-Vera et al., 2016); however, for these applications, the outcome in question (e. g. high school dropouts or freshmen attrition) is not evenly balanced between students who exhibit the outcome and those who do not. The Cross Industry Standard Process for Data Mining (CRISP-DM) process model was used for this study, providing a data mining framework and opportunities to resolve the problems identified in prior Step 1 outcomes research.

The six phases of the CRISP-DM process model (Wirth & Hipp, 2000) guided the design the design, development, and implementation of this study. This process model was selected because it has been found to be more suited for novice researchers (Kurgan

& Musilek, 2006), and because of the suggested tasks in each phase it is more likely to be adopted by medical school administrators who desire to replicate this study in the future. A detailed explanation of the tasks completed in each phase can be found in Chapter 3. A summary of major accomplishments from the process model follows.

Outcome data of medical students matriculating between 2013 and 2014 was provided by the School of Medicine at Baylor College of Medicine (BCM), a private medical school located in Houston, Texas. BCM accepts approximately 2% of applicants for each admission cycle, and is known for USMLE outcomes well above national averages. Step 1 is taken at the end of the preclinical curriculum, approximately 18 months after matriculation. Preadmission variables, such as age, gender, MCAT scores, and UGPA, final preclinical course grades, and the CBSE score were extracted from the student information system. Three business objectives and data mining goals guided the overall data mining approach: (1) use common classification data mining algorithms to determine the factors associated with Step 1 failures, (2) use preadmission variables and final course grades from the preclinical curriculum to determine the point in time which bests identifies students at risk of failing Step 1, and (3) use common data sampling methods to determine the best approach to address the expected Step 1 outcome imbalance problem.

Prior to modeling, the BCM student outcomes dataset was modified to resolve prior research problems of low sample size of Step 1 failures and the imbalance between Step 1 failed and passing observations. The number of Step 1 failed observations was increased by considering all scores within one standard deviation above the passing score of 192 as a failed score. Three data sampling methods were used to prepare the datasets

for modeling. In addition to the original unbalanced dataset, random under sampling (RUS) was used to randomly remove Step 1 passing observations to match the number of failed observations. Random oversampling (ROS) was used to replicate the failed Step 1 observations to match the number of passing observations. Finally, the synthetic minority oversampling technique (SMOTE) was used to create synthetic failed Step 1 observations, which closely resembled the actual failed observations, until the total number of failed observations matched the number of passing observations.

Eight experiments were conducted in a stepwise manner, beginning with a dataset of preadmission variables, adding final grades from the six preclinical course blocks and the final CBSE outcome in the order taken by BCM students. Four datasets, one unbalanced and three balanced, were used to train, test, and validate models created using the CART data mining algorithm. Accuracy, precision, recall, and the F-measure was noted to determine the highest performing model for each experiment. Model prediction accuracy is often used as a measure of success, but can be misleading because the majority class contributes much more to overall accuracy then the minority class (Marquez-Vera et al., 2016). For this study accuracy reflects the number of passing Step 1 observations correctly predicted. Because this is not the outcome under investigation the F-measure was used to determine the best performing model. A summary of top model performance all experiments is shown in Table 21. Full experiment results follow in the appendices.

Model effectiveness peaked with experiment 2 using a prediction model consisting of preadmission variables and final course grades for the first block of preclinical courses, dropping slightly when the second block of course grades were

added, then a sharp drop in model effectiveness for remaining experiments. Findings from this study indicate medical students at risk of failing Step 1 can be predicted as early as the end of the first block of preclinical courses, with subsequent blocks of courses adding little value to prediction models. The CART data mining algorithm was used to determine the factors associated with Step 1 failures, with the decision tree finding students with a final grade above 85.35 in the foundational sciences course (FBS) are likely to pass Step 1. Prediction models using oversampling methods outperformed models using unbalanced datasets in terms of model effectiveness.

The last phase of the CRISP-DM process model calls for a final report summarizing the results and a plan to operationalize the models used in the study. This paper serves as the summary; however, there is an opportunity to revisit prior phases to improve the predictive ability of the models prior to deployment. For example, this study employed the CART algorithm with default parameters. This algorithm was selected as it is most appropriate for continuous predictor variables, like the final course grades used as predictor variables in this study; however, additional tuning of the CART algorithm could possibly improve model performance metrics and the ability to accurately predict medical students at-risk of failing Step 1. Moreover, a different algorithm might be more appropriate as additional predictor variables are added. The data mining framework developed in this study can be used to improve USMLE Step 1 outcomes. Findings from this study will not directly address the predicted physician shortage but could allow medical school administrations to help at-risk students so all applicants who are accepted into medical education have access to all the resources necessary to achieve successful outcomes.

Appendices

Appendix A

Experiment 1 Results

This appendix contains full results for experiment 1. Model train, test, and validate performance metrics for each dataset balancing method are shown below.

Model	TP	TN	FP	FN	Accuracy	Precision	Recall	F
Unbalanced								
Train	0	164	6	0	96.5%	0.00	0.00	0.00
Test	0	164	5	0	97.0%	0.00	0.00	0.00
Validate	0	138	8	0	94.5%	0.00	0.00	0.00
RUS								
Train	5	1	1	5	50.0%	0.83	0.50	0.63
Test	4	86	1	78	53.3%	0.80	0.05	0.09
Validate	6	81	2	57	59.6%	0.75	0.10	0.17
ROS								
Train	164	147	0	17	94.8%	1.00	0.91	0.95
Test	3	144	2	20	87.0%	0.60	0.13	0.21
Validate	2	130	6	8	90.4%	0.25	0.20	0.22
SMOTE								
Train	159	146	5	18	93.0%	0.97	0.90	0.93
Test	1	147	4	17	87.6%	0.20	0.06	0.09
Validate	2	127	6	11	88.4%	0.25	0.15	0.19

Relative predictor importance and rules tree for RUS.

Predictor	Relative Importance
Total UGPA	0.69
Science UGPA	0.06
Age	0.06
PS	0.06
VR	0.06
MCAT Total	0.06

Rules Tree

- Total UGPA <= 3.885 [Mode: Fail] => Fail
- Total UGPA > 3.885 [Mode: Pass] => Pass

Relative predictor importance and rules tree for ROS.

Predictor	Relative Importance
VR	0.49
Total UGPA	0.30
BS	0.09
Age	0.07
MCAT Total	0.01
PS	0.01
Science UGPA	0.01
URM	0.01

Rules Tree

- VR <= 8.500 [Mode: Fail]
    - Total UGPA <= 3.740 [Mode: Pass] => Pass
    - Total UGPA > 3.740 [Mode: Fail] => Fail
- VR > 8.500 [Mode: Pass]
    - Total UGPA <= 3.725 [Mode: Fail]
        - BS <= 11.500 [Mode: Pass] => Pass
        - BS > 11.500 [Mode: Fail]
            - Age <= 22.500 [Mode: Fail] => Fail
            - Age > 22.500 [Mode: Pass] => Pass
    - Total UGPA > 3.725 [Mode: Pass] => Pass

Relative predictor importance and rules tree for SMOTE.

Predictor	Relative Importance
Total UGPA	0.37
VR	0.37
BS	0.10
Age	0.06
MCAT Total	0.05
PS	0.01
Gender	0.01
Science UGPA	0.01
URM	0.01

Rules Tree

- VR <= 8.998 [Mode: Fail]
    - Total UGPA <= 3.739 [Mode: Pass] => Pass
    - Total UGPA > 3.739 [Mode: Fail] => Fail
- VR > 8.998 [Mode: Pass]
    - Total UGPA <= 3.787 [Mode: Fail]
        - MCAT Total <= 35.950 [Mode: Fail]
            - BS <= 11.013 [Mode: Pass] => Pass
            - BS > 11.013 [Mode: Fail]
                - Age <= 22.500 [Mode: Fail] => Fail
                - Age > 22.500 [Mode: Pass] => Pass
        - MCAT Total > 35.950 [Mode: Pass] => Pass
    - Total UGPA > 3.787 [Mode: Pass] => Pass

Appendix B

Experiment 2 Results

This appendix contains full results for experiment 2. Model train, test, and validate performance metrics for each dataset balancing method are shown below.

Model	TP	TN	FP	FN	Accuracy	Precision	Recall	F
Unbalanced								
Train	0	164	6	0	96.5%	0.00	0.00	0.00
Test	0	164	5	0	97.0%	0.00	0.00	0.00
Validate	0	138	8	0	94.5%	0.00	0.00	0.00
RUS								
Train	5	5	1	1	83.3%	0.83	0.83	0.83
Test	4	86	1	78	53.3%	0.80	0.05	0.09
Validate	6	81	2	57	59.6%	0.75	0.10	0.17
ROS								
Train	164	159	0	5	98.5%	1.00	0.97	0.98
Test	2	159	3	5	95.3%	0.40	0.29	0.33
Validate	3	137	5	1	95.9%	0.38	0.75	0.50
SMOTE								
Train	164	153	0	11	96.6%	1.00	0.94	0.97
Test	2	160	3	4	95.9%	0.40	0.33	0.36
Validate	4	136	4	2	95.9%	0.50	0.67	0.57

Relative predictor importance and rules tree for RUS.

Predictor	Relative Importance
Total UGPA	0.6946
Science UGPA	0.0611
Block 1 FBS	0.0611
Block 1 PP1	0.0611
Age	0.0611
PS	0.0611

Rules Tree

- Total UGPA <= 3.885 [Mode: Fail] => Fail
- Total UGPA > 3.885 [Mode: Pass] => Pass

Relative predictor importance and rules tree for ROS.

Predictor	Relative Importance
Block 1 FBS	0.3722
VR	0.3666
Total UGPA	0.1405
Age	0.0746
MCAT Total	0.0077
Block 1 PP1	0.0077
PS	0.0077
BS	0.0077
Science UGPA	0.0077
URM	0.0077

Rules Tree

- Block 1 FBS <= 85.350 [Mode: Fail]
    - VR <= 8.500 [Mode: Fail] => Fail
    - VR > 8.500 [Mode: Fail]
        - Total UGPA <= 3.725 [Mode: Fail]
            - VR <= 10.500 [Mode: Pass] => Pass
            - VR > 10.500 [Mode: Fail]
                - Age <= 22.500 [Mode: Fail] => Fail
                - Age > 22.500 [Mode: Pass] => Pass
        - Total UGPA > 3.725 [Mode: Pass] => Pass
- Block 1 FBS > 85.350 [Mode: Pass] => Pass

Relative predictor importance and rules tree for SMOTE.

Predictor	Relative Importance
Block 1 FBS	0.4015
Science UGPA	0.1845
VR	0.1769
Age	0.136
BS	0.0617
PS	0.0144
MCAT Total	0.0063
Block 1 PP1	0.0063
Total UGPA	0.0063
URM	0.0063

Rules Tree

- Block 1 FBS <= 85.349 [Mode: Fail]
    - VR <= 8.998 [Mode: Fail] => Fail
    - VR > 8.998 [Mode: Fail]
        - Science UGPA <= 3.744 [Mode: Fail]
            - BS <= 11.013 [Mode: Pass] => Pass
            - BS > 11.013 [Mode: Fail]
                - Age <= 22.500 [Mode: Fail] => Fail
                - Age > 22.500 [Mode: Pass] => Pass
        - Science UGPA > 3.744 [Mode: Pass]
            - PS <= 9 [Mode: Fail] => Fail
            - PS > 9 [Mode: Pass] => Pass
- Block 1 FBS > 85.349 [Mode: Pass] => Pass

Appendix C

Experiment 3 Results

This appendix contains full results for experiment 3. Model train, test, and validate performance metrics for each dataset balancing method are shown below.

Model	TP	TN	FP	FN	Accuracy	Precision	Recall	F
Unbalanced								
Train	0	164	6	0	96.5%	0.00	0.00	0.00
Test	0	164	5	0	97.0%	0.00	0.00	0.00
Validate	0	138	8	0	94.5%	0.00	0.00	0.00
RUS								
Train	5	5	1	1	83.3%	0.83	0.83	0.83
Test	4	86	1	78	53.3%	0.80	0.05	0.09
Validate	6	81	2	57	59.6%	0.75	0.10	0.17
ROS								
Train	164	155	0	9	97.3%	1.00	0.95	0.97
Test	1	158	4	6	94.1%	0.20	0.14	0.17
Validate	3	136	5	2	95.2%	0.38	0.60	0.46
SMOTE								
Train	162	138	2	26	91.5%	0.99	0.86	0.92
Test	2	160	3	4	95.9%	0.40	0.33	0.36
Validate	8	123	0	15	89.7%	1.00	0.35	0.52

Relative predictor importance and rules tree for RUS.

Predictor	Relative Importance
Total UGPA	0.3381
Science UGPA	0.1324
Block 1 FBS	0.1324
Block 1 PP1	0.1324
Block 2 IPD	0.1324
Block 2 HNA	0.1324

Rules Tree

- Total UGPA <= 3.885 [Mode: Fail] => Fail
- Total UGPA > 3.885 [Mode: Pass] => Pass

Relative predictor importance and rules tree for ROS.

Predictor	Relative Importance
Block 2 IPD	0.5928
Block 1 PP1	0.2391
BS	0.0403
Block 1 FBS	0.016
Block 2 PHR	0.016
Total UGPA	0.016
Age	0.016
PS	0.016
MCAT Total	0.016
VR	0.016

Rules Tree

- Block 2 IPD <= 79.067 [Mode: Fail]
    - Block 1 PP1 <= 95.250 [Mode: Fail] => Fail
    - Block 1 PP1 > 95.250 [Mode: Pass] => Pass
- Block 2 IPD > 79.067 [Mode: Pass]
    - BS <= 8 [Mode: Fail] => Fail
    - BS > 8 [Mode: Pass] => Pass

Relative predictor importance and rules tree for SMOTE.

Predictor	Relative Importance
Block 2 IPD	0.9344
Block 1 FBS	0.0131
Block 2 HNA	0.0131
Block 2 PHR	0.0131
Block 1 PP1	0.0131
MCAT Total	0.0131

Rules Tree

- Block 2 IPD <= 83.451 [Mode: Fail] => Fail
- Block 2 IPD > 83.451 [Mode: Pass] => Pass

Appendix D

Experiment 4 Results

This appendix contains full results for experiment 4. Model train, test, and validate performance metrics for each dataset balancing method are shown below.

Model	TP	TN	FP	FN	Accuracy	Precision	Recall	F
Unbalanced								
Train	0	164	6	0	96.5%	0.00	0.00	0.00
Test	0	164	5	0	97.0%	0.00	0.00	0.00
Validate	0	138	8	0	94.5%	0.00	0.00	0.00
RUS								
Train	5	5	1	1	83.3%	0.83	0.83	0.83
Test	4	86	1	78	53.3%	0.80	0.05	0.09
Validate	6	81	2	57	59.6%	0.75	0.10	0.17
ROS								
Train	164	151	0	13	96.0%	1.00	0.93	0.96
Test	1	161	4	3	95.9%	0.20	0.25	0.22
Validate	1	132	7	6	91.1%	0.13	0.14	0.13
SMOTE								
Train	164	151	0	13	96.0%	1.00	0.93	0.96
Test	1	161	4	3	95.9%	0.20	0.25	0.22
Validate	1	132	7	6	91.1%	0.13	0.14	0.13

Relative predictor importance and rules tree for RUS.

Predictor	Relative Importance
Total UGPA	0.3381
Science UGPA	0.1324
Block 1 FBS	0.1324
Block 1 PP1	0.1324
Block 2 IPD	0.1324
Block 2 HNA	0.1324

Rules Tree

- Total UGPA <= 3.885 [Mode: Fail] => Fail
- Total UGPA > 3.885 [Mode: Pass] => Pass

Relative predictor importance and rules tree for ROS.

Predictor	Relative Importance
Block 3 NRS	0.9419
Block 2 IPD	0.0116
Block 1 FBS	0.0116
Block 2 HNA	0.0116
Block 2 PHR	0.0116
VR	0.0116

Rules Tree

- Block 3 NRS <= 80.050 [Mode: Fail] => Fail
- Block 3 NRS > 80.050 [Mode: Pass] => Pass

Relative predictor importance and rules tree for SMOTE.

Predictor	Relative Importance
Block 3 NRS	0.9628
Block 2 IPD	0.0074
Block 1 FBS	0.0074
Block 2 HNA	0.0074
Block 2 PHR	0.0074
MCAT Total	0.0074

Rules Tree

- Block 3 NRS <= 80.050 [Mode: Fail] => Fail
- Block 3 NRS > 80.050 [Mode: Pass] => Pass

Appendix E

Experiment 5 Results

This appendix contains full results for experiment 5. Model train, test, and validate performance metrics for each dataset balancing method are shown below.

Model	TP	TN	FP	FN	Accuracy	Precision	Recall	F
Unbalanced								
Train	0	164	6	0	96.5%	0.00	0.00	0.00
Test	0	164	5	0	97.0%	0.00	0.00	0.00
Validate	0	138	8	0	94.5%	0.00	0.00	0.00
RUS								
Train	5	5	1	1	83.3%	0.83	0.83	0.83
Test	4	86	1	78	53.3%	0.80	0.05	0.09
Validate	6	81	2	57	59.6%	0.75	0.10	0.17
ROS								
Train	164	151	0	13	96.0%	1.00	0.93	0.96
Test	1	161	4	3	95.9%	0.20	0.25	0.22
Validate	1	132	7	6	91.1%	0.13	0.14	0.13
SMOTE								
Train	164	151	0	13	96.0%	1.00	0.93	0.96
Test	1	161	4	3	95.9%	0.20	0.25	0.22
Validate	1	132	7	6	91.1%	0.13	0.14	0.13

Relative predictor importance and rules tree for RUS.

Predictor	Relative Importance
Total UGPA	0.3381
Science UGPA	0.1324
Block 1 FBS	0.1324
Block 1 PP1	0.1324
Block 2 IPD	0.1324
Block 2 HNA	0.1324

Rules Tree

- Total UGPA <= 3.885 [Mode: Fail] => Fail
- Total UGPA > 3.885 [Mode: Pass] => Pass

Relative predictor importance and rules tree for ROS.

Predictor	Relative Importance
Block 3 NRS	0.9456
Block 4 BES	0.0109
Block 2 IPD	0.0109
Block 1 FBS	0.0109
Block 4 IND	0.0109
Block 2 HNA	0.0109

Rules Tree

- Block 3 NRS <= 80.050 [Mode: Fail] => Fail
- Block 3 NRS > 80.050 [Mode: Pass] => Pass

Relative predictor importance and rules tree for SMOTE.

Predictor	Relative Importance
Block 3 NRS	0.9517
Block 2 IPD	0.0097
Block 4 BES	0.0097
Block 4 IND	0.0097
Block 1 FBS	0.0097
Block 2 HNA	0.0097

Rules Tree

- Block 3 NRS <= 80.050 [Mode: Fail] => Fail
- Block 3 NRS > 80.050 [Mode: Pass] => Pass

Appendix F

Experiment 6 Results

This appendix contains full results for experiment 6. Model train, test, and validate performance metrics for each dataset balancing method are shown below.

Model	TP	TN	FP	FN	Accuracy	Precision	Recall	F
Unbalanced								
Train	0	164	6	0	96.5%	0.00	0.00	0.00
Test	0	164	5	0	97.0%	0.00	0.00	0.00
Validate	0	138	8	0	94.5%	0.00	0.00	0.00
RUS								
Train	5	5	1	1	83.3%	0.83	0.83	0.83
Test	4	86	1	78	53.3%	0.80	0.05	0.09
Validate	6	81	2	57	59.6%	0.75	0.10	0.17
ROS								
Train	164	151	0	13	96.0%	1.00	0.93	0.96
Test	1	161	4	3	95.9%	0.20	0.25	0.22
Validate	1	132	7	6	91.1%	0.13	0.14	0.13
SMOTE								
Train	164	159	0	5	98.5%	1.00	0.97	0.98
Test	0	163	5	1	96.4%	0.00	0.00	0.00
Validate	1	136	7	2	93.8%	0.13	0.33	0.18

Relative predictor importance and rules tree for RUS.

Predictor	Relative Importance
Total UGPA	0.3381
Science UGPA	0.1324
Block 1 FBS	0.1324
Block 1 PP1	0.1324
Block 2 IPD	0.1324
Block 2 HNA	0.1324

Rules Tree

- Total UGPA <= 3.885 [Mode: Fail] => Fail
- Total UGPA > 3.885 [Mode: Pass] => Pass

Relative predictor importance and rules tree for ROS.

Predictor	Relative Importance
Block 3 NRS	0.9456
Block 4 BES	0.0109
Block 2 IPD	0.0109
Block 1 FBS	0.0109
Block 4 IND	0.0109
Block 2 HNA	0.0109

Rules Tree

- Block 3 NRS <= 80.050 [Mode: Fail] => Fail
- Block 3 NRS > 80.050 [Mode: Pass] => Pass

Relative predictor importance and rules tree for SMOTE.

Predictor	Relative Importance
Block 3 NRS	0.7171
Block 5 ETH	0.2226
Block 2 IPD	0.0075
Block 4 BES	0.0075
Block 4 IND	0.0075
Block 1 FBS	0.0075
Block 2 HNA	0.0075
Block 1 PP1	0.0075
Block 2 PHR	0.0075
Block 5 PP2	0.0075

Rules Tree

- Block 3 NRS <= 80.050 [Mode: Fail]
    - Block 3 NRS <= 74.550 [Mode: Pass] => Pass
    - Block 3 NRS > 74.550 [Mode: Fail]
        - Block 5 ETH <= 86.550 [Mode: Fail] => Fail
        - Block 5 ETH > 86.550 [Mode: Pass] => Pass
- Block 3 NRS > 80.050 [Mode: Pass] => Pass

Appendix G

Experiment 7 Results

This appendix contains full results for experiment 7. Model train, test, and validate performance metrics for each dataset balancing method are shown below.

Model	TP	TN	FP	FN	Accuracy	Precision	Recall	F
Unbalanced								
Train	0	164	6	0	96.5%	0.00	0.00	0.00
Test	0	164	5	0	97.0%	0.00	0.00	0.00
Validate	0	138	8	0	94.5%	0.00	0.00	0.00
RUS								
Train	5	5	1	1	83.3%	0.83	0.83	0.83
Test	4	86	1	78	53.3%	0.80	0.05	0.09
Validate	6	81	2	57	59.6%	0.75	0.10	0.17
ROS								
Train	164	151	0	13	96.0%	1.00	0.93	0.96
Test	1	161	4	3	95.9%	0.20	0.25	0.22
Validate	1	132	7	6	91.1%	0.13	0.14	0.13
SMOTE								
Train	164	159	0	5	98.5%	1.00	0.97	0.98
Test	1	163	4	1	97.0%	0.20	0.50	0.29
Validate	1	136	7	2	93.8%	0.13	0.33	0.18

Relative predictor importance and rules tree for RUS.

Predictor	Relative Importance
Total UGPA	0.3381
Science UGPA	0.1324
Block 1 FBS	0.1324
Block 1 PP1	0.1324
Block 2 IPD	0.1324
Block 2 HNA	0.1324

Rules Tree

- Total UGPA <= 3.885 [Mode: Fail] => Fail
- Total UGPA > 3.885 [Mode: Pass] => Pass

Relative predictor importance and rules tree for ROS.

Predictor	Relative Importance
Block 3 NRS	0.9363
Block 6 CAR	0.0127
Block 6 PP3	0.0127
Block 4 BES	0.0127
Block 2 IPD	0.0127
Block 6 RNL	0.0127

Rules Tree

- Block 3 NRS <= 80.050 [Mode: Fail] => Fail
- Block 3 NRS > 80.050 [Mode: Pass] => Pass

Relative predictor importance and rules tree for SMOTE.

Predictor	Relative Importance
Block 3 NRS	0.5324
Block 6 PP3	0.4021
Block 2 IPD	0.0082
Block 4 BES	0.0082
Block 6 CAR	0.0082
Block 6 RNL	0.0082
Block 5 ETH	0.0082
Block 6 GST	0.0082
Block 1 FBS	0.0082
Block 2 HNA	0.0082

Rules Tree

- Block 3 NRS <= 80.050 [Mode: Fail]
    - Block 6 PP3 <= 89.300 [Mode: Fail] => Fail
    - Block 6 PP3 > 89.300 [Mode: Pass] => Pass
- Block 3 NRS > 80.050 [Mode: Pass] => Pass

Appendix H

Experiment 8 Results

This appendix contains full results for experiment 8. Model train, test, and validate performance metrics for each dataset balancing method are shown below.

Model	TP	TN	FP	FN	Accuracy	Precision	Recall	F
Unbalanced								
Train	0	164	6	0	96.5%	0.00	0.00	0.00
Test	0	164	5	0	97.0%	0.00	0.00	0.00
Validate	0	138	8	0	94.5%	0.00	0.00	0.00
RUS								
Train	5	5	1	1	83.3%	0.83	0.83	0.83
Test	4	86	1	78	53.3%	0.80	0.05	0.09
Validate	6	81	2	57	59.6%	0.75	0.10	0.17
ROS								
Train	164	151	0	13	96.0%	1.00	0.93	0.96
Test	1	161	4	3	95.9%	0.20	0.25	0.22
Validate	1	132	7	6	91.1%	0.13	0.14	0.13
SMOTE								
Train	164	159	0	5	98.5%	1.00	0.97	0.98
Test	1	163	4	1	97.0%	0.20	0.50	0.29
Validate	1	136	7	2	93.8%	0.13	0.33	0.18

Relative predictor importance and rules tree for RUS.

Predictor	Relative Importance
Total UGPA	0.3381
Science UGPA	0.1324
Block 1 FBS	0.1324
Block 1 PP1	0.1324
Block 2 IPD	0.1324
Block 2 HNA	0.1324

Rules Tree

- Total UGPA <= 3.885 [Mode: Fail] => Fail
- Total UGPA > 3.885 [Mode: Pass] => Pass

Relative predictor importance and rules tree for ROS.

Predictor	Relative Importance
Block 3 NRS	0.9363
Block 6 CAR	0.0127
Block 6 PP3	0.0127
Block 4 BES	0.0127
Block 2 IPD	0.0127
Block 6 RNL	0.0127

Rules Tree

- Block 3 NRS <= 80.050 [Mode: Fail] => Fail
- Block 3 NRS > 80.050 [Mode: Pass] => Pass

Relative predictor importance and rules tree for SMOTE.

Predictor	Relative Importance
Block 3 NRS	0.5008
Block 6 PP3	0.4308
Block 2 IPD	0.0098
Block 4 BES	0.0098
Block 6 CAR	0.0098
Block 6 CBSE	0.0098
Block 5 ETH	0.0098
Block 6 GST	0.0098
Block 1 FBS	0.0098

Rules Tree

- Block 3 NRS <= 80.050 [Mode: Fail]
    - Block 6 PP3 <= 89.300 [Mode: Fail] => Fail
    - Block 6 PP3 > 89.300 [Mode: Pass] => Pass
- Block 3 NRS > 80.050 [Mode: Pass] => Pass

References

Abeysinghe, W., Hung, C. C., Bechikh, S., Wang, X., & Rattani, A. (2018, October). Clustering algorithms on imbalanced data using the SMOTE technique for image segmentation. In *Proceedings of the 2018 Conference on Research in Adaptive and Convergent Systems* (pp. 17-22). ACM.

About the AAMC. (n.d.). Retrieved from https://www.aamc.org/about

About the MCAT Exam. (n.d.). Retrieved from https://students-residents.aamc.org/applying-medical-school/taking-mcat-exam/about-mcat-exam/

About the NBME (2018). Retrieved from https://www.nbme.org/about/index.html.

Admissions frequently asked questions. (n.d.). Retrieved from https://www.bcm.edu/education/schools/medical-school/md-program/admissions/faqs.

Alsaffar, A. H. (2017). Empirical study on the effect of using synthetic attributes on classification algorithms. *International Journal of Intelligent Computing and Cybernetics*, *10*(2), 111-129.

Andriole, D. A., & Jeffe, D. B. (2010). Prematriculation variables associated with suboptimal outcomes for the 1994-1999 cohort of US medical school matriculants. *JAMA*, *304*(11), 1212–1219. https://doi.org/10.1001/jama.2010.1321

Arlot, S., & Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistics Surveys*, 4, 40-79.

Barber, C., Hammond, R., Gula, L., Tithecott, G., & Chahine, S. (2018). In search of black swans: identifying students at risk of failing licensing examinations. *Academic Medicine*, *93*(3), 478-485.

Baylor makes leap in U.S. News & World Report rankings. (2018, March 20). Retrieved from https://www.bcm.edu/news/awards-honors-college/baylor-us-news-world-report-rankings.

Becker, S. A., Brown, M., Dahlstrom, E., Davis, A., DePaul, K., Diaz, V., & Pomerantz, J. (2018). NMC Horizon Report: 2018 Higher Education Edition. Louisville, CO: EDUCAUSE.

Branco, P., Torgo, L., & Ribeiro, R. P. (2016). A survey of predictive modeling on imbalanced domains. *ACM Computing Surveys (CSUR)*, *49*(2), 31-50.

Burk-Rafel, J., Santen, S. A., & Purkiss, J. (2017). Study behaviors and USMLE Step 1

performance: Implications of a student self-directed parallel curriculum. *Academic Medicine*, *92*(11S), S67-S74.

Burish, M. J., Fredericks, C. A., Engstrom, J. W., Tateo, V. L., & Josephson, S. A. (2015). Predicting success: what medical student measures predict resident performance in neurology? *Clinical Neurology and Neurosurgery*, *135*, 69-72.

Burns, E. R., & Garrett, J. (2015). Student failures on first-year medical basic science courses and the USMLE step 1: A retrospective study over a 20-year period. *Anatomical Sciences Education*, *8*(2), 120–125. https://doi.org/10.1002/ase.1462

Capers, Q., McDougle, L., & Clinchot, D. M. (2018). Strategies for Achieving Diversity through Medical School Admissions. *Journal of Health Care For the Poor and Underserved*, *29*(1), 9-18.

Chauhan, R., Kaur, H., & Sharma, S. (2016, August). A Feature Based Approach for Medical Databases. In *Proceedings of the International Conference on Advances in Information Communication Technology & Computing* (p. 94). ACM.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321-357.

Chen, L.-Y. A., & Fawcett, T. N. (2016). Using data mining strategies in clinical decision making. *CIN: Computers, Informatics, Nursing*, *34*(10), 448–454. https://doi.org/10.1097/CIN.0000000000000282

Chen, W., Xie, X., Wang, J., Pradhan, B., Hong, H., Bui, D. T., ... & Ma, J. (2017). A comparative study of logistic model tree, random forest, and classification and regression tree models for spatial prediction of landslide susceptibility. *Catena*, *151*, 147-160.

Clow, D. (2013). An overview of learning analytics. *Teaching in Higher Education*, *18*(6), 683–695. https://doi.org/10.1080/13562517.2013.827653

Conrad, S., Addams, A., & Young, G. (2016). Holistic review in medical school admissions and selection: A strategic, mission-driven response to shifting societal needs. *Academic Medicine*, *91*(11), 1472–1474. https://doi.org/10.1097/ACM.0000000000001403

Coumarbatch, J., Robinson, L., Thomas, R., & Bridge, P. D. (2010). Strategies for identifying students at risk for USMLE step 1 failure. *Family Medicine*, *42*(2), 105–110.

Dakduk, S., Malavé, J., Torres, C. C., Montesinos, H., & Michelena, L. (2016). Admission Criteria for MBA Programs: A Review. *SAGE Open*, *6*(4),

2158244016669395.

Dall, T., West, T., Chakrabarti, R., Reynolds, R., & Iacobucci, W. (2018). The complexities of physician supply and demand: Projections from 2016 to 2030. Retrieved from https://aamc-black.global.ssl.fastly.net/production/media/filer_public/85/d7/85d7b689-f417-4ef0-97fb-ecc129836829/aamc_2018_workforce_projections_update_april_11_2018.pdf

Delen, D. (2010). A comparative analysis of machine learning techniques for student retention management. *Decision Support Systems, 49*(4), 498-506.

DeZee, K. J., Artino, A. R., Elnicki, D. M., Hemmer, P. a., & Durning, S. J. (2012). Medical education in the United States of America. *Medical Teacher*, *34*(June), 521–525. https://doi.org/10.3109/0142159X.2012.668248

Donnon, T., Paolucci, E. O., & Violato, C. (2007). The predictive validity of the MCAT for medical school performance and medical board licensing examinations: a meta-analysis of the published research. *Academic Medicine*, *82*(1), 100–106. https://doi.org/10.1097/01.ACM.0000249878.25186.b7

Dunleavy, D. M., Kroopnick, M. H., Dowd, K. W., Searcy, C. A., & Zhao, X. (2013). The predictive validity of the MCAT exam in relation to academic performance through medical school. *Academic Medicine*, *88*(5), 666–671. https://doi.org/10.1097/ACM.0b013e3182864299

Educational Data Mining (n.d.) Retrieved from http://www.educationaldatamining.org.

Eisen, D. B., Schupp, C. W., Isseroff, R. R., Ibrahimi, O. A., Ledo, L., & Armstrong, A. W. (2015). Does class attendance matter? Results from a second-year medical school dermatology cohort study. *International Journal of Dermatology*, *54*(7), 807-816.

Ellaway, R. H., Malhi, R., Bajaj, S., Walker, I., & Myhre, D. (2018). A critical scoping review of the connections between social mission and medical school admissions: BEME Guide No. 47. *Medical Teacher*, *40*(3), 219-226.

Ellaway, R. H., Pusic, M. V., Galbraith, R. M., & Cameron, T. (2014). Developing the role of big data and analytics in health professional education. *Medical Teacher, 36*(3). 216-222.

Elks, M. L., Herbert-Carter, J., Smith, M., Klement, B., Knight, B. B., & Anachebe, N. F. (2018). Shifting the curve: Fostering academic success in a diverse student body. *Academic Medicine*, *93*(1), 66-70.

Fernández, A., Garcia, S., Herrera, F., & Chawla, N. V. (2018). SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year

Anniversary. *Journal of Artificial Intelligence Research, 61*, 863-905.

Gay, S. E., Santen, S. A., Mangrulkar, R. S., Sisson, T. H., Ross, P. T., & Zaidi, N. L. B. (2018). The influence of MCAT and GPA preadmission academic metrics on interview scores. *Advances in Health Sciences Education, 23*(1), 151-158.

Gauer, J. L., & Jackson, J. B. (2018). Relationships of demographic variables to USMLE physician licensing exam scores: A statistical analysis on five years of medical student data. *Advances In Medical Education and Practice*, *9*, 39-44.

Gauer, J. L., Wolff, J. M., & Brooks Jackson, J. (2016). Do MCAT scores predict USMLE scores? An analysis on 5 years of medical student data. *Medical Education Online*, *21*(1). https://doi.org/10.3402/meo.v21.31795

Giordano, C., Hutchinson, D., & Peppler, R. (2016). A Predictive Model for USMLE Step 1 Scores. *Cureus*, *8*(9), 1–8. https://doi.org/10.7759/cureus.769

Glaros, A. G., Hanson, A., & Adkison, L. R. (2014). Early prediction of medical student performance on initial licensing examinations. *Medical Science Educator*, *24*(3), 291–295. https://doi.org/10.1007/s40670-014-0053-y

Gohara, S., Shapiro, J. I., Jacob, A. N., Khuder, S. A., Gandy, R. A., Metting, P. J., … Kleshinski, and J. (2011). Joining the Conversation: Predictors of Success on the United States Medical Licensing Examinations (USMLE). *Learning Assistance Review*, *16*(1), 11–20. Retrieved from http://www.nclca.org/tlar.html

Gray, C. C. & Perkins, D. (2019). Utilizing early engagement and machine learning to predict student outcomes. *Computers & Education, 131*, 22-32.

Gullo, C. A., McCarthy, M. J., Shapiro, J. I., & Miller, B. L. (2015). Predicting medical student success on licensure exams. *Medical Science Educator*, *25*(4), 447-453.

Gullo, C. A. (2016). The future is in the numbers: the power of predictive analysis in the biomedical educational environment. *Medical Education Online*, *21*(1), 32516.

Hairrell, A. R., Smith, S., McIntosh, D., & Chico, D. E. (2016). Impact of pre-matriculation instruction on student acculturation and first-year academic performance in medical school. *Medical Science Educator, 26*(4), 519-523.

Heck, A. J., Gibbons, L., Ketter, S. J., Furlano, A., & Prest, L. (2017). A Survey of the Design of Pre-matriculation Courses at US Medical Schools. *Medical Science Educator*, *27*(2), 229-236.

Herzog, S. (2006). Estimating student retention and degree-completion time: Decision trees and neural networks vis-à-vis regression. *New Directions for Institutional Research, 2006*(131), 17-33.

Holistic Review (2018). Holistic review in admissions. Retrieved from https://www.aamc.org/initiatives/holisticreview/

Hu, Y., Martindale, J. R., LeGallo, R. D., White, C. B., McGahren, E. D., & Schroen, A. T. (2016). Relationships between preclinical course grades and standardized exam performance. *Advances in Health Sciences Education*, *21*(2), 389–399. https://doi.org/10.1007/s10459-015-9637-6

Hutt, S., Gardener, M., Kamentz, D., Duckworth, A. L., & D'Mello, S. K. (2018, March). Prospectively predicting 4-year college graduation from student applications. In *Proceedings of the 8th International Conference on Learning Analytics and Knowledge* (pp. 280-289). ACM.

Ilin, D. A., & Krivtsov, V. E. (2015). Creating training datasets for OCR in Mobile Device Video Stream. In Proceedings from the 29th European Conference on Modelling and Simulation (pp. 516-520).

Johnson, L., Smith, R., Willis, H., Levine, A., and Haywood, K., (2011). *The 2011 Horizon Report.* Austin, Texas: The New Media Consortium.

Johnson, L., Adams, S., and Cummins, M. (2012). The *NMC Horizon Report: 2012 Higher Education Edition.* Austin, Texas: The New Media Consortium.

Johnson, L., Adams Becker, S., Cummins, M., Estrada, V., Freeman, A., and Ludgate, H. (2013). *NMC Horizon Report: 2013 Higher Education Edition*. Austin, Texas: The New Media Consortium.

Johnson, L., Adams Becker, S., Estrada, V., Freeman, A. (2014). *NMC Horizon Report: 2014 Higher Education Edition*. Austin, Texas: The New Media Consortium.

Johnson, L., Adams Becker, S., Estrada, V., and Freeman, A. (2015). *NMC Horizon Report: 2015 Higher Education Edition*. Austin, Texas: The New Media Consortium.

Jones, R. F., & Thomae-Forgues, M. (1984). Validity of the MCAT in predicting performance in the first two years of medical school. *Academic Medicine*, *59*(6), 455–464.

Journal of Educational Data Mining (n.d.). Retrieved from https://jedm.educationaldatamining.org.

Journal of Learning Analytics (n.d.). Retrieved from https://learning-analytics.info/journals/index.php/jla.

Julian, E. R. (2005). Validity of the Medical College Admission Test for predicting medical school performance. *Academic Medicine*, *80*(10), 910–917.

https://doi.org/10.1097/00001888-200510000-00010

Khalil, M. K., Hawkins, H. G., Crespo, L. M., & Buggy, J. (2017). The design and development of prediction models for maximizing students' academic achievement. *Medical Science Educator*, 1-7.

Kleshinski, J., Khuder, S. A., Shapiro, J. I., & Gold, J. P. (2009). Impact of preadmission variables on USMLE step 1 and step 2 performance. *Advances in Health Sciences Education*, *14*(1), 69–78. https://doi.org/10.1007/s10459-007-9087-x

Klieger, D. M., Bridgeman, B., Tannenbaum, R. J., Cline, F. A., & Olivera-Aguilar, M. (2018). The Validity of GRE General Test Scores for Predicting Academic Performance at US Law Schools. ETS Research Report Series.

Klieger, D. M., Cline, F. A., Holtzman, S. L., Minsky, J. L., & Lorenz, F. (2014). New perspectives on the validity of the GRE General Test for predicting graduate school grades. *ETS Research Report Series*, *2014*(2), 1-62.

Kostopoulos, G., & Lipitakis, A. (2017). Predicting student performance in distance higher education using active aearning. In *International Conference on Engineering Applications of Neural Networks* (pp. 75–86). Springer International. https://doi.org/10.1007/978-3-319-65172-9

Kurgan, L. A., & Musilek, P. (2006). A survey of knowledge discovery and data mining process models. *The Knowledge Engineering Review*, *21*(1), 1-24.

Lakkaraju, H., Aguiar, E., Shan, C., Miller, D., Bhanpuri, N., Ghani, R., & Addison, K. L. (2015, August). A machine learning framework to identify students at risk of adverse academic outcomes. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1909-1918). ACM.

Lauría, E. J., Baron, J. D., Devireddy, M., Sundararaju, V., & Jayaprakash, S. M. (2012, April). Mining academic data to improve college student retention: An open source perspective. In Proceedings of the 2nd International Conference on Learning Analytics and Knowledge (pp. 139-142). ACM.

Lee, M. W., Johnson, T. R., & Kibble, J. (2017). Development of statistical models to predict medical student performance on the USMLE Step 1 as a catalyst for deployment of student services. *Medical Science Educator*, 1–9. https://doi.org/10.1007/s40670-017-0452-y

Lee, E. B., Kim, J., & Lee, S. G. (2017). Predicting customer churn in mobile industry using data mining technology. *Industrial Management & Data Systems, 117*(1), 90-109.

Lieberman, S. A., Frye, A. W., Thomas, L., Rabek, J. P., & Anderson, G. D. (2008). Comprehensive changes in the learning environment: subsequent step 1 scores of academically at-risk students. *Academic Medicine*, *83*(10 Suppl), S49-52. https://doi.org/10.1097/ACM.0b013e318183e2d0

Márquez-Vera, C., Cano, A., Romero, C., Noaman, A. Y. M., Mousa Fardoun, H., & Ventura, S. (2016). Early dropout prediction using data mining: A case study with high school students. *Expert Systems*, *33*(1), 107–124. https://doi.org/10.1111/exsy.12135

Monroe, A., Quinn, E., Samuelson, W., Dunleavy, D. M., & Dowd, K. W. (2013). An overview of the medical school admission process and use of applicant data in decision making. *Academic Medicine*, *88*(5), 672–681. https://doi.org/10.1097/ACM.0b013e31828bf252

Nagasawa, D. T., Beckett, J. S., Lagman, C., Chung, L. K., Schmidt, B., Safaee, M., ... & Yang, I. (2017). United States Medical Licensing Examination Step 1 scores directly correlate with American Board of Neurological Surgery scores: a single-institution experience. *World Neurosurgery*, *98*, 427-431.

National Resident Matching Program Data Release and Research Committee. (2018). *Results of the 2018 NRMP program director survey*. Washington, DC. Retrieved from https://www.nrmp.org/wp-content/uploads/2018/07/NRMP-2018-Program-Director-Survey-for-WWW.pdf

Navigate your journey from pre-med to residency. (n.d.). Retrieved March 21, 2018, from https://students-residents.aamc.org/

Record of success. (n.d.). Retrieved from https://media.bcm.edu/documents/2017/55/bcm-detail-enrollment-2017-2018.pdf.

Rumsfeld, J. S., Joynt, K. E., & Maddox, T. M. (2016). Big data analytics to improve cardiovascular care: promise and challenges. *Nature Reviews Cardiology*, *13*(6), 350–359. https://doi.org/10.1038/nrcardio.2016.42

Saguil, A., Dong, T., Gingerich, R. J., Swygert, K., LaRochelle, J. S., Artino Jr, A. R., ... & Durning, S. J. (2015). Does the MCAT predict medical school and PGY-1 performance?. *Military Medicine*, *180*(suppl_4), 4-11.

Schneid, S. D., Apperson, A., Laiken, N., Mandel, J., Kelly, C. J., & Brandl, K. (2018). A summer prematriculation program to help students succeed in medical school. *Advances in Health Sciences Education*, 1-13.

Schwartz, L. F., Lineberry, M., Park, Y. S., Kamin, C. S., & Hyderi, A. A. (2018). Development and Evaluation of a Student-Initiated Test Preparation Program for the

USMLE Step 1 Examination. *Teaching and Learning in Medicine*, *30*(2), 193-201.

Segal, S. S., Giordani, B., Gillum, L. H., & Johnson, N. (1999). The Academic Support Program at the University of Michigan School of Medicine. *Academic Medicine*, *74*(4), 383–385. Retrieved from http://journals.lww.com/academicmedicine/Fulltext/1999/04000/The_Academic_Support_Program_at_the_University_of.31.aspx

Siemens, G. (2013). Learning analytics: The emergence of a discipline. *American Behavioral Scientist, 57*(10), 1380-1400.

Siemens, G., & Baker, R. S. (2012, April). Learning analytics and educational data mining: Towards communication and collaboration. In Proceedings of the 2nd International Conference on Learning Analytics and Knowledge (pp. 252-254). ACM.

Sesate, D. B., Milem, J. F., McIntosh, K. L., & Bryan, W. P. (2017). Coupling admissions and curricular data to predict medical student outcomes. *Research in Higher Education*, *58*(3), 295–312. https://doi.org/10.1007/s11162-016-9426-y

Slater, S., Joksimović, S., Kovanovic, V., Baker, R. S., & Gasevic, D. (2017). Tools for educational data mining: A review. *Journal of Educational and Behavioral Statistics*, *42*(1), 85-106.

Stratton, T. D., & Elam, C. L. (2014). A holistic review of the medical school admission process: Examining correlates of academic underperformance. *Medical Education Online*, *19*(1), 22919.

Student enrollment statistics. (2017, September 19). Retrieved from https://media.bcm.edu/documents/2017/55/bcm-detail-enrollment-2017-2018.pdf.

Thammasiri, D., Delen, D., Meesad, P., & Kasap, N. (2014). A critical assessment of imbalanced class distribution problem: The case of predicting freshmen student attrition. *Expert Systems with Applications*, *41*(2), 321–330. https://doi.org/10.1016/j.eswa.2013.07.046

Underrepresented in medicine education. (2018). Retrieved from https://www.aamc.org/initiatives/urm/

USMLE. (2018). USMLE Overview. Retrieved from: https://www.usmle.org/bulletin/overview/

Wei, X., Jiang, F., Wei, F., Zhang, J., Liao, W., & Cheng, S. (2017, May). An ensemble model for diabetes diagnosis in large-scale and imbalanced dataset. In *Proceedings of the Computing Frontiers Conference* (pp. 71-78). ACM.

Wiley, A., & Koenig, J. A. (1996). The validity of the Medical College Admission Test for predicting performance in the first two years of medical school. *Academic Medicine*, *71*(10 Suppl), S83–S85.

Winston, K. a, van der Vleuten, C. P. M., & Scherpbier, A. J. J. a. (2014). Prediction and prevention of failure: An early intervention to assist at-risk medical students. *Medical Teacher*, *36*(1), 25–31. https://doi.org/10.3109/0142159X.2013.836270

Wirth, R., & Hipp, J. (2000). CRISP-DM: Towards a standard process model for data mining. In *Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining* (pp. 29–39). Manchester, UK. https://doi.org/10.1.1.198.5133

Zhao, S., King, I., Lyu, M. R., Zeng, J., & Yuan, M. (2017, August). Mining business opportunities from location-based social networks. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 1037-1040). ACM.